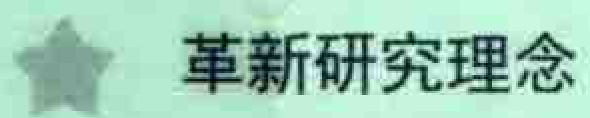


格致方法·定量研究系列　吴晓刚　主编

因素调查实验

[德] 卡特琳·奥斯普格（Katrin Auspurg）
托马斯·欣茨（Thomas Hinz）著
陈霜叶 荣佳妮 译　叶晓阳 王哲 校



革新研究理念



丰富研究工具

最权威、最前沿的定量研究方法指南

70

格致出版社　上海人民出版社



因素调查实验是一种在问卷调查中进行实验的社会科学研究方法。这种方法适用于识别与评估研究对象对真实世界的信念、感受、社会偏好、决策与评价原则的社会科学研究，例如人口流动性问题、收入公平问题、社会歧视问题、住房偏好、移民选择与教育政策评估等，在社会学各领域应用广泛，具有巨大前景。本书是该领域的入门级书籍，两位作者基于自身实验设计的经验，深入浅出地介绍了核心概念、相关研究及注意事项。

主要特点

- 提供了“菜谱”式的指导和具体步骤演示
- 每章都提供了方法概要与具体案例，适合研究生与有一定量化方法基础的研究者

您可以通过如下方式联系到我们：
邮箱：hibooks@hibooks.cn


微信


天猫



定价：45.00元
易文网：www.ewen.co
格致网：www.hibooks.cn

格致方法·定量研究系列　吴晓刚　主编

因素调查实验

[德] 卡特琳·奥斯普格（Katrin Auspurg）
托马斯·欣茨（Thomas Hinz）　著
陈霜叶　荣佳妮　译
叶晓阳　王　哲　校

SAGE Publications ,Inc.
格致出版社　上海人民出版社

图书在版编目(CIP)数据

因素调查实验/(德)卡特琳·奥斯普格,(德)托马斯·欣茨著;陈霜叶,荣佳妮译.—上海:格致出版社:上海人民出版社,2021.3
(格致方法.定量研究系列)
ISBN 978-7-5432-3227-3

Ⅰ.①因… Ⅱ.①卡… ②托… ③陈… ④荣… Ⅲ.①社会科学-研究方法 Ⅳ.①C3

中国版本图书馆CIP数据核字(2021)第026322号

责任编辑 贺俊逸

格致方法·定量研究系列
因素调查实验
[德]卡特琳·奥斯普格 托马斯·欣茨 著
陈霜叶 荣佳妮 译
叶晓阳 王 哲 校

出　　版 格致出版社
上海人民出版社
(200001 上海福建中路193号)
发　　行 上海人民出版社发行中心
印　　刷 浙江临安曙光印务有限公司
开　　本 920×1168 1/32
印　　张 6.5
字　　数 129,000
版　　次 2021年3月第1版
印　　次 2021年3月第1次印刷
ISBN 978-7-5432-3227-3/C·251
定　　价 45.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的"格致方法·定量研究系列"丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于2011年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于2012年以单行本的形式再次出版发行，共37本。我们衷心感谢广大读者的支持和建议。

随着与SAGE出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”（Statistics for Social Science）为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课（事实上，大部分硕士生在修完第一门课后都会继续选修第二门课）。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复，但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂，与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在"格致方法 · 定量研究系列"丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办"应用社会科学研究方法研修班"，至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

从根本上区分观测数据和实验数据是基础统计学课程的标准做法。从观测数据中进行因果推断非常困难，而从随机比较实验中进行因果推断就相对直观。

社会科学和行为科学使用的观测数据多来自抽样调查，而心理学或社会心理学研究多使用实验数据，数据通常来自少数大学生志愿者。这样的实验有明显的局限性，尤其是在如何将实验结果推广到不同于参与当前研究的其他群体之外的推广性问题方面。

然而，研究者可以在样本调查中进行随机比较实验，为参与研究的被试提供有效的因果推断[有时称为"内在效度"，1966 年坎贝尔和斯坦利(Campbell & Stanley)在颇有影响力的专著中引入了这一术语]，同时促进对总体的一般推广性(外在效度)。大多数的问卷实验研究是相对简单的，只使用少量的(通常只有两个)实验条件，而且常常是为了发现问卷设计存在的问题而设计的，例如，选项措辞选择(alternative question wording)的影响。康弗斯和普雷瑟(Converse &

Presser, 1986)在Sage出版的《问卷调查问题》中，描述了几个类似的“分签—投票”(split-ballot)实验。[*] 相比之下，卢维埃(Louviere, 1988)在论述“指标联合分析”(metric conjoint analysis)的书中描述了多因素实验设计，设计了多个同时影响的实验条件，用来研究顾客的决策。虽然这种方法主要应用于市场营销研究中，但也可以追溯到20世纪50年代初彼得·罗西(Peter Rossi)采用的“虚拟情境法”(vignette)[**] 在问卷调查研究应用的传统。“虚拟情境法”是通过建构一段描述或虚构故事的方式，同时改变两个或多个实验因素。这一方法也称为“因素调查设计”或“因素调查实验”(Jasso, 2006; Rossi, 1951)。

以费希尔(R.A. Fisher, 1925) 20世纪初对随机实验的贡献为基础，实验设计在统计学中是一项严肃的主题。此后，费希尔和其他统计学家在这一课题上发展了大量文献，逐步形成了研究若干同时产生影响的且相互作用的潜在因素的有效设计。虽然实验研究对于心理学家而言是很常见的，但是其他社会和行为科学家并不经常使用。在《因素调查实验》中，卡特琳·奥斯普格(Katrin Auspurg)和托马斯·欣茨(Thomas Hinz)基于实验设计的统计学文献，向读者介绍了核心概念，例如“随机区组(块)”(blocking)和“部分混淆”(partial confounding)，并演示这些技术在社会调查实验

* 这是指把一群研究对象分成两个或者两个以上的群体，以稍有不同的问卷来搜集信息的方法。——译者注

** 这是该方法的核心概念与方法的精髓之处，也是一个难翻成中文的概念。Vignette的字典意思是指一段能清晰展示人物特征和局势的短文。这里指根据实验需要设置的一段情境描述或者虚拟情境，有的地方翻译为“情境法”。我们刻意选择“虚拟情境”来突出这种虚拟性。——译者注

操作中的应用。他们同时还详细描述了调查实验中所产生的问题。我希望这本书能够进一步促进调查实验方法的发展,并有助于提升因素调查研究的质量、深度和广度。

约翰·福克斯

参考文献

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Dallas, TX: Houghton Mifflin.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Newbury Park, CA: Sage.

Fisher, R. A. (1925). *Statistical methods for research workers*. New York: Hafner.

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods and Research*, 34(3), 334-423.

Louviere, J. J. (1988). *Analyzing decision making: Metric conjoint analysis.* Newbury Park, CA: Sage.

Rossi, P. H. (1951). *The application of latent structure analysis to the study of social stratification*. PhD dissertation, Columbia University.

前 言

这本书不长，但历史却很久。我们在 2008 年就开始了关于因素调查的研究。我们都对在问卷调查中实施实验饶有兴趣。我们的研究小组在地区流动性问题、收入公平性问题、社会歧视问题、住房偏好问题、移民问题以及教育政策的评估问题等许多项目中都运用了因素调查。我们深信，这个工具在试图识别人们的决策和评价原则以及检验因果推断假设的所有社会学科领域中存在巨大价值。几年前，许多研究人员还不知道因素调查，但现在因素调查在问卷调查研究和许多大规模的家庭调查中已经是一个公认的研究工具。当我们计划进行关于因素调查方法论方面的研究时，可以明显发现以往的研究主要集中关注一些实质性的问题，而在建立和实施因素调查研究实验以及分析结果数据中，对各种重要的方法论问题的关注并不多。虽然早在 20 世纪七八十年代就有关于方法论问题的文献，但这些文献对于当前研究者是陌生的，并且显然没有考虑到当前软件或实验设计的最新发展。这个项目是我们与比勒费尔德大学（Bielefeld University）的斯特凡·利比希（Stefan Liebig）和卡斯滕·绍尔

(Carsten Sauer)合作的，是德国研究基金会在调查方法领域的优先项目(DFG)，它解决了广泛的方法论问题，例如对虚拟情境的抽样、回答量表、处理不合逻辑的虚拟情境、呈现方式、多水平的数据分析等。在项目的进程中，我们受邀参加许多课程讲座和研讨会，并经常被问及额外的关于方法论的建议。这样的经历和动机正是本书的灵感来源：将我们在因素调查设计方面的专业知识尽可能全面地汇编成册，并以“菜谱”的形式呈现。到目前为止，实施因素调查的难度相当高。研究人员必须结合不同领域的知识，包括实验设计、调查研究和多水平分析。此外，他们常常需要找到自己的途径来应用这一方法。我们希望这本书能使得克服这些困难变得更容易些。除了提供因素调查实验的“菜谱”外，我们还向读者介绍类似方法的最新进展，例如联合分析、选择实验和一些更先进的统计工具。为了更好地达到我们的目的，本书附带了在线资料，其中有示例数据和大量有用的(统计)文档，演示了如何操作不同的步骤。

在本书的编写过程中，我们得到了许多人的支持。首先，我们要感谢因素调查和公正研究的先驱之一的吉列米娜·杰索(Guillermina Jasso)。她的热情帮助是我们持续努力的决定性因素。她与彼得·罗西合作的作品是我们最大的灵感来源；她还帮助我们在旧金山、亚特兰大和拉斯维加斯举办的美国社会学会年会上，组织了关于因素调查方法的国际讨论。所有参与讨论的人都间接为本书做出了贡献。其次，是来自比勒费尔德大学的斯特凡·利比希，他是该活动的联合组织者之一，也是在德国推广因素调查实验的学者之一。多年来，我们在由诺曼·布朗(Norman Braun)、安德

烈亚斯·迪克曼(Andreas Diekmann)和约瑟夫·布鲁德尔(Josef Bruderl)在威尼斯国际大学主持的"理性选择"研讨会上展示了本书的部分内容。我们感谢他们给予的机会,并感谢他们的评论。特别要感谢卡尔·迪特尔·奥普(Karl Dieter Opp)的支持和帮助,因为他首先在各种研究项目中使用因素调查方法。

本书的文本是以回环的形式呈现。通过约翰·福克斯组织的同行评议以及他提供的深思熟虑的评论和优秀的建议,我们受益匪浅。两位匿名的评审提供了富有建设性的反馈。此外,我们还邀请一些学者对早期版本进行评论。所有的工作都极大地改进了文稿。我们要感谢汤姆·斯尼德斯(Tom Snijders)、托马斯·奥古斯丁(Thomas Augustin)、本·詹恩(Ben Jann)、卡斯滕·绍尔、克里斯蒂安·博佐扬(Christiane Bozoyan)和索尼娅·波因特纳(Sonja Pointner)的帮助,以及康斯坦茨大学几位研究助理从学生的角度提供的有用反馈。詹姆斯·迪斯利(James Disley)的语言编辑使我们的想法能更清晰地传递给读者。最后,是海伦·萨蒙(Helen Salmon)自身杰出的出版经验、鼓励性的反馈和她的管理技巧,保证了项目的持续发展。感谢所有人。如有任何的错误都是我们的责任。

许多正在进行的研究活动将有助于进一步阐释因素调查实验及其相关方法,我们将通过更新在线资料来跟进。我们将非常感谢读者在这方面提供的任何额外的信息。

卡特琳·奥斯普格和托马斯·欣茨

分别于康斯坦茨和法兰克福,2014 年 6 月

目 录



绪　论

为了回应保罗·F.拉扎斯菲尔德(Paul F.Lazarsfeld)提出的一个想法,彼得·H.罗西在他1951年的毕业论文中引入了“因素调查”(factorial survey, FS)方法。在这篇论文中,罗西应用了因素调查方法来测量家庭的社会地位(Rossi, 1979)。受访者被要求在9分量表中判断一些虚构的家庭地位(Rossi, Sampson, Bose, Jasso, & Passel, 1974)。这些虚构家庭被设定为由不同职业和不同层次教育水平的丈夫和妻子所组成。代表不同家庭的职业和教育特征的数值有系统性的变化,从而可以估计实验变量对受访者对家庭地位等级评定的影响。在这一设计中,罗西和他的同事也评估了受访者对地位等级评定中单一维度重要性的认可程度(Rossi et al., 1974)。研究者进而对不同社会人群(如不同年龄或受教育程度)对社会地位和分层的感知是一致或分化得出结论。

罗西的方法开启了更一般性的想法,即开发一种技术来评估社会规范、态度和定义背后的判断原则。规范性的判断和定义(例如,社会地位等级评定)通常基于多方面的属性,而非单一维度。这些属性进而被整合到一个单一且一致的判断中。例如,对犯罪者的公平判决涉及若干方面,诸如犯罪发生的环境、犯罪实施者的特点以及对受害者的损害等,

这需要在考虑相互作用的基础上进行评估(例如,参见 J.L. Miller, Rossi, & Simpson, 1986)。同样地,公平的福利金和税率的概念(Liebig & Mau, 2005)以及性骚扰和虐待儿童的定义(Garret, 1982; O'Toole, Webster, O'Toole, & Lucal, 1999)也是基于多个维度。与多维度方法相一致的是,因素调查方法背后的动机是向受访者呈现类似于真实世界评估的刺激,使他们在多个维度之间进行权衡。相对于使用单一项目的问题,这种方法能够更精确地确定评估背后的判断原则。用更技术性的方式表达则是,因素调查的核心特征是在一项调查中实施多维度的实验设计。它要求受访者对"虚拟情境"进行判断。这种"虚拟情境"就是对假设的环境、物或人的多维度描述。这些维度的价值(层次)在不同的虚拟情境中进行随机分配,从而可以评估这些维度的价值(层次)对受访者判断的影响。因素调查方法是诸多调查实验方法中的一种(Mutz, 2011)。不同于经典的"分签—投票"实验(在方法研究中主要用于改进回答量表或题目类型),使用因素调查方法的调查研究使用多因子设计,且关注实质性的问题,例如受访者的态度或行为意图。通常只涉及单一变量或是相当狭隘的方法论内容的经典"分签—投票"实验不在本书探讨的范围之内。感兴趣的读者可以参考调查研究的具体文献(Groves et al., 2009; Sniderman & Grob, 1996)。

因素调查方法的应用范围不仅包括之前提到的规范性判断和定义,还包括有意图的行动(例如迁移或不迁移、雇用或不雇用一个工作申请者的意图:Abraham, Auspurg, & Hinz, 2010; de Wolf & van der Velden, 2001)。因素调查的测量目标还包括受访者对真实世界(相信某事是真实存在

的）的信念、感受和想法。随着越来越多的大规模家庭调查中使用因素调查方法，以及国际核心期刊发表越来越多的相关论文，因素调查方法的重要性日益显著（概述参见 Mutz, 2011; Wallander, 2009）。

此书对何时、为什么以及如何使用因素调查方法提供了操作性的介绍。相对于有限的若干因素调查方法论的出版物（Alexander & Becker, 1978; Dülmer, 2007; Hox, Kreft, & Hermkens, 1991; Jasso, 2006; Rossi & Anderson, 1982），这本书提供了从最初的想法到创建一个因素调查研究和数据分析的统计技术所需要的各个实际步骤的全面介绍。高阶使用者将会受益于单个章节中对更复杂问题的处理（例如当构造一个“虚拟情境”样本时如何处理不合逻辑的情况）。本书中也囊括了调查研究和计算机技术的最新进展（例如用于建立实验设计的软件工具）。许多步骤都是通过实例来说明的（www.sagepub.com/auspurg_hinz 提供了如数据文件等补充材料；这个网页还提供了与因素调查技术相关的技术术语表）。

本书结构如下：第 2 章详细介绍了因素调查方法使用的范围、优点和应用，还介绍了主要的技术术语。第 3 章说明了如何设计一个因素调查研究实验，包括选择维度、水平和它们的组合方式。本章尤其关注能够以最大的精确度和统计效力进行因果推断的情境样本构建的实验设计理论。此外，本章还提供了“虚拟情境”和受访者的最适样本量。第 4 章介绍了如何将“虚拟情境”纳入调查，包括创建文本情境和选择适当的被试样本、回答量表、调查模式以及田野调查的准备细节（例如，前测）。第 5 章呈现了数据分析技术，包括

数据的准备、多层数据结构的分析方法(例如,稳健标准误和多水平分析)和因素调查研究中可使用的特殊测量方法(例如,支付意愿的估计)。第 6 章比较了因素调查方法与其他多因素实验调查方法(联合分析与选择实验法),以帮助读者从一系列相似的技术中进行有效选择。同时,本章还讨论了效度问题。第 7 章包含了对方法论最新进展的评论(例如,将因素调查与定性访问技术的结合),以及对最相关的研究的总结性建议。尽管第 2 章可能会增加读者对整本书中使用的技术术语的熟悉程度,但高阶读者可以直接进入后面的相关章节。

第2章

为何以及何时使用因素调查分析

第 1 节 | 以收入公平研究为例

假设你要研究影响对员工收入是否公平看法的相关因素，那么这些决定人们对公平工资看法的标准主要与工作绩效(比如工作努力程度或教育)相关吗？还是这些标准主要与需求(比如家庭中孩子的数量)相关？男性员工和女性员工的收入应该一样吗？如果不是，什么样的工资差距才算是公平的？对于这些问题，在不同社会群体之间是否达成了共识，还是存在判断原则上的分歧？

调查这些问题的标准方法是简单地询问受访者有关其态度的单项问题，或是用李克特量表(Likert scale)提出一系列问题。以公平收入为例，人们可能直接问，公平收入的标准应该是什么，或者年轻者与年长者、受教育水平较低者与较高者、男性员工与女性员工是否应该得到同等报酬？可是，这些都是非常抽象的问题，而且强有力的社会和法律规范已经规定了男女平等的待遇，包括同工同酬。因此，我们不清楚通过直接提问是否可以获得受访者的真实看法，还是只能获得那些社会认同的观点。此外，通过这种方法很难解释男性和女性员工应该获得不同报酬的回答。受访者认为女性应该获得更高的工资还是更低的工资？在多大程度上？这样的陈述是否表达了一种歧视的偏好(即支持同工不同酬)，或者受

访者是否认为男女员工在劳动力市场中具有不同的特征(如不同的工作时间),以此来证明不平等报酬的合理性?

由于上述原因,我们希望能够更深入地了解人们判断和决策的原则,而不仅仅使用单项问题。与单项问题相比,多维度的情境化描述会导致更精细的提问。因此,受访者回答更不容易受到社会倾向性偏见的影响(Alexander & Becker, 1978; Auspurg, Hinz, Sauer, & Liebig, 2014)。此外,对情境更为详细的描述使得情境刺激更为标准化,为受访者判断原则提供更深入的理解。

如上所述,收入公平性评价是因素调查方法最突出的应用之一(例如,参见 Alves & Rossi, 1978; Hermkens & Boerman, 1989; Jasso & Rossi, 1977; Jasso & Webster, 1997, 1999; Shepelak & Alwin, 1986)。关于这个话题的示例情境如图 2.1 所示。在实验中,以下维度在不同的虚拟情境中随机变化(楷体标注):年龄、性别、教育、孩子数量、职业、工作经验、任期、月收入。[1]一般而言,受访者会评估多个"虚拟情境"(常见的数量是 10—20 个)。在理想的情况下,

---

一名接受过职业培训的 35 岁男子有 4 个孩子。
他是一名电气工程师,几乎没什么工作经验。

他曾在公司工作过很短的一段时间。
他每月的总收入达到 2 500 欧元(税前收入)。

从你的角度看,这个人的月收入是公平的?还是太低?或太高?

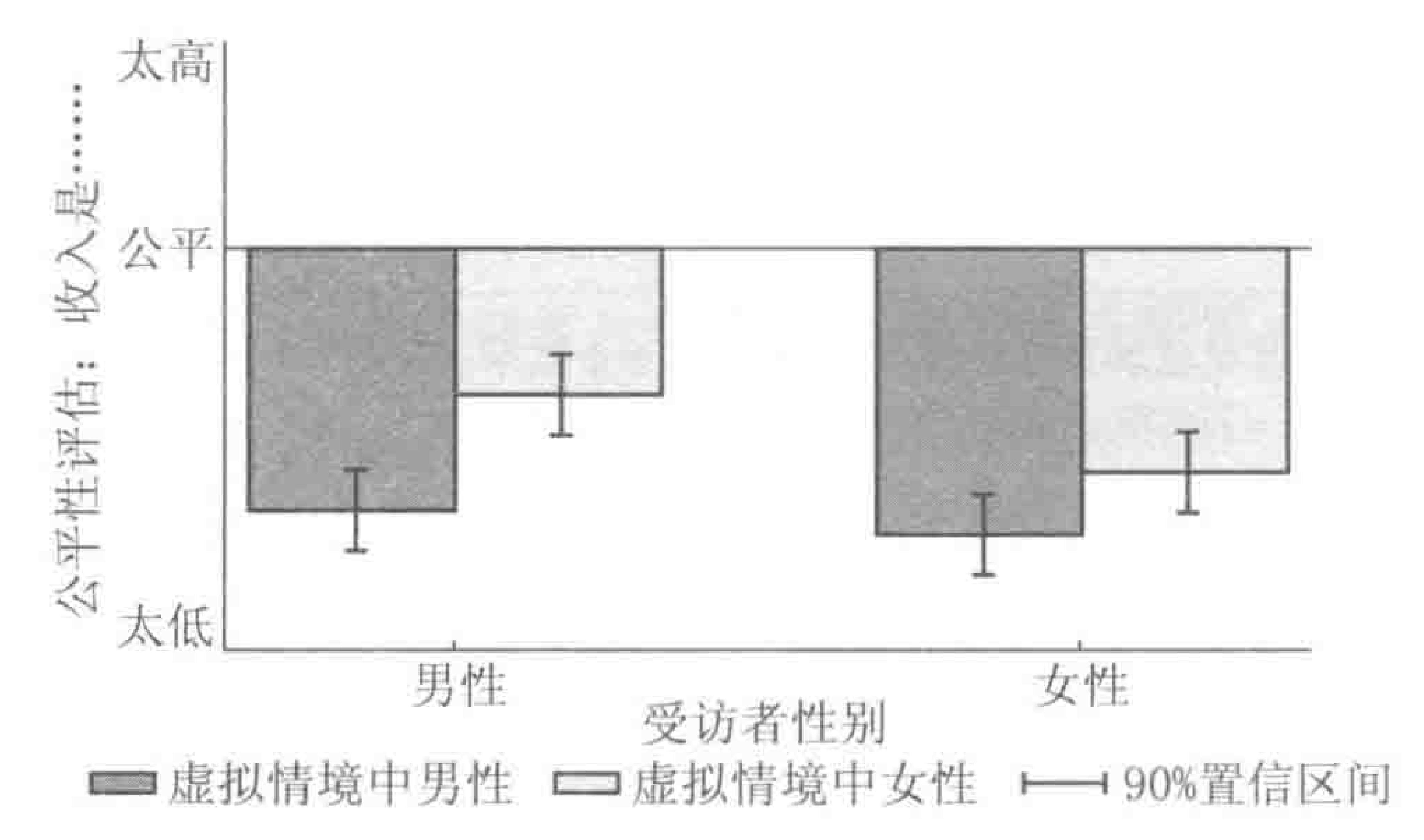
不公平 太低					公平					不公平 太高
−5	−4	−3	−2	−1	0	1	2	3	4	5
●	●	●	●	●	●	●	●	●	●	●

资料来源:Justice of Earnings Survey Konstanz, 2008。

图 2.1　关于收入公平的因素调查的示例情境

所有维度是完全独立的(即它们之间是不相关的)。基于这个条件以及向受访者随机呈现不同"虚拟情境",不同情境水平下发现的评分的系统性差异直接揭示了这些维度的影响。例如,对年轻者和年长者或男性和女性的情境特征直接展示了年龄或性别维度对收入公平性评价的因果影响。

图 2.2 显示了一项针对德国成年人的调查中性别维度的结果(详情参阅 Auspurg et al., 2014)。平均而言,虚拟情境中男性的收入更容易被评价为过低。对于男性受访者而言,这一差异在统计学上是显著的(深色条形表示虚拟的男性性别特征,顶端的钉形表示置信区间,负值表示其报酬被评价为过低)。[2]换句话说,男性受访者特别支持"公平的性别收入差距"(justice gender pay gap, JGPG),表明男性员工的收入应该高于女性员工的收入。因为男性和女性的情境特征平均而言在维度和水平上是相同的(即所有情境总和),所以评价的差异并不是其他情境特征(例如年龄或教育)所导致的。



注:n=4 426 个虚拟情境和 n=445 个调查对象。均值估计和置信区间是对男性和女性调查者在控制虚拟个体收入之后的随机斜率回归而分别估计的。

资料来源:Justice of Earnings Survey Konstanz, 2008。

图 2.2 虚拟情境中的人物(VP)性别与受访者在 90% 置信区间(CIs)的平均值估计

因素调查方法也可以用来评估不同维度的影响和它们确切的(货币化的)权衡取舍(trade-offs)。这些评估可以通过来自受访者 j 的单个情境 i 下的公平性评价 Y_{ij} 对 p 个情境维度(如年龄或性别,用 X 表示)的回归分析实现。这个线性回归方程在 2.1 方程中公式化,用 β(β_1 到 β_p)表示回归系数,ε_{ij} 表示判断中的随机误差,n_d 表示对单个受访者虚拟情境呈现的情境数量,n_r 表示受访者的数量:

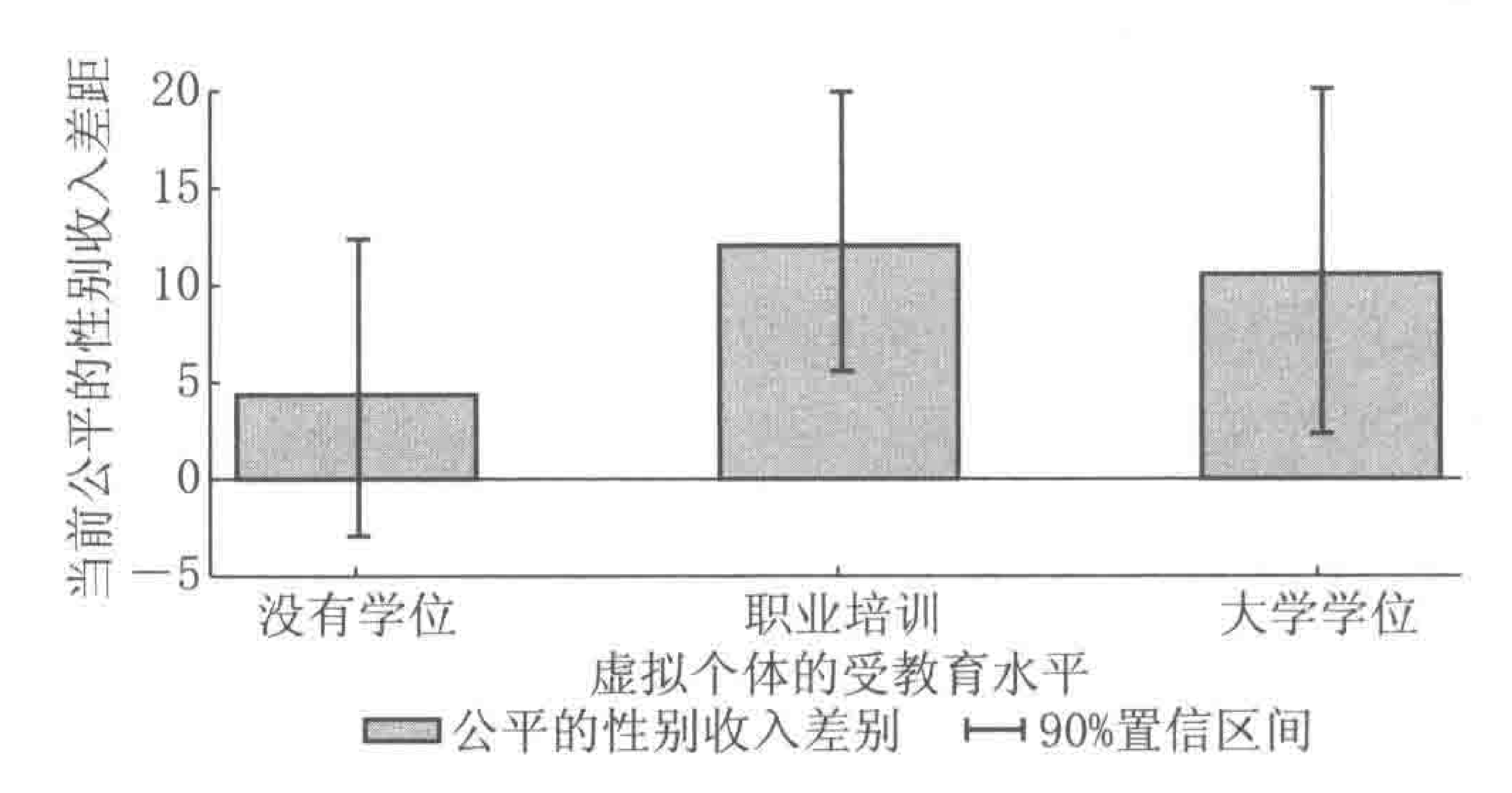
$$Y_{ij}=\beta_0+\beta_1 X_{ij1}+\beta_2 X_{ij2}+\cdots+\beta_p X_{ijp}+\varepsilon_{ij}$$

$$\text{其中},i=1,\ \cdots,\ n_d;\ j=1,\ \cdots,\ n_r。\qquad [2.1]$$

回归系数表示对虚拟情境中单个维度(如年龄或性别)对受访者对情境的判断的线性可加影响。要评估公平的性别收入差距,就必须估算出在受访者的判断中可以抵消性别特征带来的差异的收入水平。在统计学上,这个过程需要性别变量的收入弹性(详见第 5 章)。[3]接下来将会进一步使用回归方程对虚拟情境维度在受访者评估的影响进行建模。

多维的方法能够使研究者更多地了解产生不公平判断的机制。例如,公平支付差距是否会随员工的教育水平变化?要回答这一问题,我们只需要对不同虚拟情境组合下的判断原则进行分别估计,或者将交互项添加到回归方程中(详见第 5 章)。图 2.3 显示了一项在德国使用因素调查方法的研究结果。具体而言,公平的性别收入差距对于教育水平较高的人(即拥有大学学位或接受过职业培训)尤为明显。受访者似乎只支持了在基本工资中很小的性别差异,但是在教育回报率上性别差距很大(男性被分配更高的教育回报


率）。在美国也发现了类似的结果（Jasso & Webster，1997）。



注：n=4 426 个虚拟情境和 n=445 个调查对象。适用 Stata 的 *wtp* 计算得出的平均收入差距和置信区间，在计算不同教育水平的调查者的随机回归系数后的分别估计。

图 2.3 虚拟个体的受教育水平与公平的性别收入差距，90%置信区间

同样地，确定受访者判断原则的一致性程度也是可能的。图 2.2 给出了对这一现象的初步影响，它对比了男性和女性受访者平均分值。亚群体差异和社会共识程度通常是社会科学领域（如社会学或政治学）最感兴趣的议题。使用因素调查方法的研究实例包括研究犯罪严重性的感知共识的形成（Rossi et al.，1974）以及证明福利支付合法性的特征（Will，1993；关于该方法在衡量亚群体差异的潜力的一般方法学讨论参见 Byers & Zeller，1998）。类似地，因素调查规划行动或行为意图中的（特定于亚群体的）标准，例如，雇用具有不同维度特征的员工的意愿或者接受不同属性的工作机会的意愿（例如，见 Abraham et al.，2010；de Wolf & van der Velden，2001）。在著名期刊上发表的大量研究都说明了使用因素调查设计可以研究的问题的广泛性（概述详见 Mutz，2011；Wallander，2009；在第 2 章第 3 节也能看到）。

小结

因素调查方法旨在更深入地了解受访者的判断原则。因素调查的核心元素是多维度的实验设计。受访者对刺激情境进行判断，即对假设场景或对象的描述（虚拟情境）。在这些虚拟情境中，特征的层次（维度）的变化是系统性的。因素调查方法的另一个重要方面是向受访者随机分配虚拟情境。因素调查方法既可以用来识别社会群体共享的判断原则，也可以确定亚群体的差异。

第2节 | 实验在调查中的优势

调查研究和实验研究相结合是因素调查设计的关键特征。从调查和实验中得出的结论通常是基于完全不同的逻辑形式。在调查数据的基础上,研究者通常使用多变量分析技术,通过对伪相关的事后控制来建立对研究对象随机抽样的因果推断(即他们试图控制所有可能会得出错误的因果结论的异质性)。相比之下,实验控制了一个刺激因素,并将足够多的参与者随机分配到不同的实验条件下(即存在或者不存在刺激的情境)。对不同条件下结果的比较并不需要多变量控制策略。平均而言,随机化还可以中和参与者"未被观察到的异质性"的影响。因此,实验具有较高的内部效度(即他们测量的是他们想要测量的东西)。干预效应可以同一种直接的方式予以解释,即通过对刺激因素的控制来排除引起结果变量变化的绝大多数的替代解释。由于实验室实验是在人为的、被动的环境中进行的,而且参与者的群体相当有限且同质化程度高(例如,主要是心理学和经济学的大学生以及其他高度自我选择的群体;Henrich, Heine, & Norenzayan, 2010),因此实验室实验缺乏外部有效性。由于不可能招募到足够数量的参与者进行多因素实验,大多数实验室研究只能测试几个实验因素。有限的因子数量损害了这类

研究的推广性和外部效度；如果缺乏刺激因素可能减弱所研究的因果关系的条件信息，则会威胁到实验的内部效度。为什么调查和实验的结合有吸引力呢？这种吸引力在于通过将实验整合到调查中，既避免了单维刺激的缺点，又增强了内部效度。与此同时，大量随机人群中招募因素调查实验的参与者也增强了外部效度（Mutz，2011）。

因素调查的研究设计显然遵循了实验研究的指导原则。研究者有意地控制刺激或干预（不同水平的维度），那么结果（即受访者的评估）与这种控制之间存在因果关系。对于任何一个受访者内部比较的设计，即每个受访者使用几个虚拟情境来了解个体的判断原则，这种优势是显而易见的。然而，在受访者之间的设计中进行受访者比较也具有明显优势。在经典的实验中，随机分配刺激因素是至关重要的。在因素调查设计中的等效因素是将不同虚拟情境与受访者进行随机匹配。如果研究者运用一种有效的设计来创造刺激因素（见第3章），那么就能以较高的精度来评估维度对个体判断的真实影响，而不存在虚假关联造成的偏差。由于这种精确性和无偏性，因素调查的方法比收集调查数据或观察数据等维度之间高度相关的方法更具有吸引力。以对公平收入评价为例，在现实中，相关维度配对，例如职业与教育，或职业与性别是相关的；因此，很难区分每个维度的单个影响。假设的虚拟情境提供了一种巧妙的方式来避免这种频繁出现的多元共线性的问题；这些维度被设计为不相关的（Rossi & Anderson，1982）。

在因素调查研究中拓展现实的能力也是有用的，因为评估（情境）的场景或被试者的假设性描述可以提供现实中（还

未）不存在的刺激因素。例如，如果要评估（假设的）政治措施或者对犯罪的适当与否的（假设的）制裁问题，那么因素调查的方法就很有意义。再例如，通过向受访者提供多种情景组合形式，因素调查可以评估无法交易的物品的支付意愿。此外，假设性的刺激因素为很少能观察到的事件（例如双职工家庭的长途搬迁）提供了研究可能（Abraham et al.，2010；Rossi & Anderson，1982）。在这些情况下，因素调查提出了关于受访者自身假设行为的问题。显然，研究者必须意识到假设和实际行动间可能存在的差距（见第 6 章）。

与经典的基于项目的调查问题相比，因素调查强制受访者在权衡取舍中作出判断。例如，如果受访者依次被问到对影响收入的特征进行排序的问题，他们可以向所有个体特征分配相同或相似的（高）重要性（Krosnick & Alwin，1988）。根据设计，因素调查强制受访者考虑可能的权衡；他们必须同时对所有维度进行评估。因此，与经典的基于项目的调查相比，在因素调查中描述的情境与现实生活更相似，因为它包括了更多复杂的需要权衡取舍的考虑。将收入和工作时间等按等距的维度变量整合为有意义的单位，如货币或工时等，使研究者能够直接计算出货币和时间等单位在这些维度上的权衡。

如前所述，因素调查另一个优势是避免社会期望偏差（social desirability bias）。如果评估任务涉及敏感性问题（例如在公平收入研究例子里的歧视问题），因素调查就不容易激起回应者选择社会认可的答案的偏误（即社会接受性偏误）。类似地，这种优势根植于多维度的评估性任务。在最近的一项研究中，受访者似乎对公平收入表达出更诚实的感

受；与单项问题相比，他们在回答因素调查时出较少表现出的社会所期望的反应（Auspurg et al.，2014）。

另一个使用因素调查的重要动机在于，如果实验设计能得到正确的实施，内部和外部效度都会提高（见第 3、4 章）。如前面提到的，内部效度表明观测到的结果变量的变化实际上是由实验刺激产生的，而外部效度则表明研究结果可以被推广到其他环境、样本或方法上。随机受访者样本是将调查研究结果推广到更广泛的受访者抽样框架内的人群中的理想选择。可是，随机抽样往往无法实现，或者说只能通过较高的调查成本实现。因素调查实验逻辑的优势是可以利用非随机样本得出关于因果机制的一般结论。此外，在较大的（随机或非随机）普通人群样本中，在不同地理区域的调查中，或者在特殊的（随机或非随机）专业人员（如社会工作者或护士）样本的调查中，因素调查可以很容易实现（Mutz，2011；Wallander，2009）。因此，因素调查的样本可以很容易地扩展到心理学和经济学大学生这种实验室调查的“标准参与者”之外的群体。总而言之，因素调查方法有助于克服众多实验研究中依赖高度特定个体的方便抽样的限制，从而提高了因素调查结果在更多异质性受访者样本中的普适性（Mutz，2011；Rossi & Anderson，1982）。回到收入公平的研究例子中，在人口样本中运用因素调查，研究者可以对比不同调查对象群体（例如不同年龄、教育背景、职业或性别的调查对象）对公平的看法，从而评估群体观念的共识程度（Jasso & Webster，1997）。这些亚群体的研究结果通常只有在随机抽取的人口样本情况下，才能够推广到抽样的总体。然而，因素调查方法可以使一个非随机的人口样本增加

结果的推广性，因为与大多数实验室研究相比，它是在更广范围的参与者中研究维度对结果变量的影响。

最后，因素调查有助于优化研究资源的使用。在许多应用中，研究对象可以对多个虚拟情境进行评估；有些研究里虚拟情境数量超过 100 个。因此，研究者在不增加招募新参与者的额外成本下可以增加观测量。这种方法更适用于多因素实验，从而可以对实验因素之间的相互作用进行评估。在一项因素调查研究中可以很容易地对数百个实验条件进行测试，这在实验室实验中是不可能的。此外，随着观测数据的增加，干预效果的估计精度也得到了提高。然而，每个对象的多种评估也会导致标准误的膨胀（见第 3 章第 5 节第二部分），进而导致精度的降低。我们在第 3 章中会更详细地讨论理想的样本数和受访者数量。因素调查方法的局限性将在第 6 章讨论。

小结

因素调查集合了实验研究和调查研究的双重优点。实验保证了内部效度，即受访者对随机分配的虚拟情境的反应只体现实验刺激（情境）的差异。调查研究又可以很容易地应用到相对异质的人群中，从而增加了结果的普适性，使结果可以在更广泛的参与者和环境中得到更多的推广，进而增加外部效度。此外，假设性的描述能克服现实中高度相关的维度问题，并对现实中很少或从未出现的情况进行考察。与其他直接提问技术相比，因素调查研究减少了社会期望偏差。多维度的设计允许进行权衡取舍的估计。

第3节 应用领域

正如最近一篇研究述评(Wallander, 2009)所总结的,因素调查已经被广泛运用于不同的学术研究和非学术领域,包括社会学、经济学、法学、心理学、政治学、护理学和市场营销研究。在社会科学内部,因素调查所涉及的主题相当广泛,有些侧重于社会规范和价值观的内容和实证影响。特别是,研究者对激活社会规范的条件感兴趣(Beck & Opp, 2001; Diefenbach & Opp, 2007; Horne, 2003; Jasso & Opp, 1997)。因素调查早期应用于测量个体和家庭的地位和威望(Meudell, 1982; Nock, 1982; Rossi, 1979)。在前文用过的对公平收入的评估研究领域里也长期使用因素调查方法(Alves & Rossi, 1978; Hermkens & Boerman, 1989; Jasso & Webster, 1997, 1999; Shepelak & Alwin, 1986)。其他的因素调查研究包括社会政策问题和正义评估,例如贫困的维度(Will, 1993)、福利支付标准或公平税率(Liebig & Mau, 2005),以及公司裁员可接受的标准(Struck, Krause, & Pfeifer, 2008)。因素调查方法也被用于性骚扰定义的研究(Garret, 1982; O'Toole et al., 1999; Rossi & Anderson, 1982)、罪犯的适当判决(Berk & Rossi, 1977; Hembroff, 1987; J. L. Miller et al., 1986)、选择移民的标准(Jasso,

1988)，以及选择医疗的标准(Hechter，Ranger-Moore，Jasso，& Horne，1999)。此外，社会歧视研究(Jann，2005；Jonh & Bates，1990)、信任的社会嵌入(Buskens & Weesie，2000)，以及家庭社会学(Abraham et al.，2010)也使用因素调查方法。穆茨(Mutz，2011)和沃兰德(Wallander，2009)的研究示范了因素调查在城市社会学和体育社会学的进一步应用。在其他研究领域，例如护理、社会工作、法律和商业，因素调查常用于研究专业人员使用的判断原则(Charlton，2002；Ludwick et al.，2004；Ludwick & Zeller，2001；Wallander，2012；Wason，Polonsky，& Hyman，2002)。

在许多社会科学研究的应用中，结果变量测量了调查对象对于遵守规范或不遵守规范行为的态度或者公正的评价(例如涉及收入、税率以及刑事处罚的严重程度)。一些研究分析了行为意图，例如，接受一份工作邀请的可能性(Abraham et al.，2010)，或从一个具有特定特征的经销商那里购买一辆二手车(例如，Buskens & Weesie，2000)。其他的例子有经济学研究中测量住房偏好(Shlay，1985)，或确定工作特征在拒绝或接受工作邀约时的决定性作用(例如，Abraham et al.，2013)。还有研究目的是调查定义(也被称为“对世界的积极信念”)，例如对性骚扰的定义或一个家庭的高社会地位的定义(例如，O'Toole et al.，1999)。

因素调查应用的调查对象与应用领域一样多元化：调查对象包括专家、专业人员、学生和普通人群等。如上所述，在常规问卷中设计实验的方式就可以招募到如社会工作者、医生、法官这些很少被招募到实验室进行实验的专业人士。

必须澄清的是，并非所有使用情境模拟的调查研究都是因素调查。在没有任何实验条件变化的情况下，虚拟情境有时会被用来激发受访者对标准化的类型情境的想象（例如，参见 Finch，1987）。在调查研究中，另一个不应与因素调查方法混淆的是锚定情境法（anchoring vigettnes approach）。在这种方法中，虚拟情境是研究者会向调查对象呈现一种假设性描述，以帮助他们标准化自己对于自身处境的主观评价。锚定情境法经常出现在对个人幸福、健康和生活满意度的研究中。一个典型的锚定情境描述了一个带有一定特征的标准化的第三者。在对第三者的情况进行评价后，调查对象会被要求对自己的情况进行评价。如果适当运用锚定情境法，那么情境就为比较调查对象的主观评价提供了一个参考框架。在不同文化中经常存在不同评价参考点，这种方法特别有助于规范不同文化背景下的评价差异。这本书并没有涉及锚定情境法（详见 King，Murrary，Salomon，& Tandon，2004；King & Wand，2007）。与因素调查相比，锚定情境法和其他形式的情境方法并不依托于多因素实验设计。因此，为了区分这两种方法，当我们用情境作为实验变量时，我们涉及的是因素调查方法（而不是其他情境方法）。

小结

在社会学、政治学、法学、护理等社会科学研究领域中，因素调查的实证研究应用很多。结果变量是激活和接受社会规范的态度或条件。一个经典的应用是对收入公平性的

评价。因素调查可以分解正义研究中的不同判断原则，并可以衡量社会规范接受的背景。此外，因素调查还能检视行为意图和定义。需要注意的是，并不是所有使用虚拟情境的研究都是因素调查。因素调查具有若干维度水平层次的实验变异和向受访者随机呈现虚拟情境的特征。

第3章

实验设计

之前的章节中介绍了因素调查研究的主要概念。从本质上说，研究者向研究对象呈现简短的描述(虚拟情境)，每个虚拟情境在水平层次上单一的属性(维度)在实验中是不同的。在实验文献中，这种维度被称为因子。

另一个重要的术语是"全情境"[虚拟情境全集(vignette universe)，也称为总体情境(vignette population)；实验文献中称为全部的因子]，它是由所有可能结合的维度构成的。全情境 N_u 的大小是所有维度和水平数量的笛卡尔乘积。例如，如果有三个维度两个水平层次、两个维度三个水平层次，以及一个维度五个水平层次，那么全集包含了 $2\times2\times2\times3\times3\times5=2^3\times3^2\times5^1=360$ 个虚拟情境。通常，这样的多因子实验设计被简单地缩写为 $2^3 3^2 5^1$ 设计，全部的数量代表了不同维度呈现的水平总数，而指数代表了涵盖这些水平的维度数。

很明显的是，全集的大小会随着维度数和水平数的增加而呈指数式增长。因此，虽然有的研究可以呈现全部的因子设计(例如，一个拥有 2^3 因子只包含了 8 个虚拟情境)，但是全情境通常包括比一个研究对象能见到的更多的虚拟情境。因此，我们必须使用虚拟情境的样本。抽样可能使全集中的

一些理想特性丢失。在全集中，所有维度和维度间的交互作用是不相关的。用更专业的术语表达，所有因子和交互性都是完全“正交”的。对于描绘一组虚拟情境样本的结果是，不再存在维度水平的所有组合，并会导致维度间的关联。在某些水平上的过度采样可能会导致单一维度呈现出不平衡的水平数量（在全集中，所有维度的水平都以相同的频率发生），这意味着维度方面的方差变小。所有这些问题都表明，实验的某些强度在丧失。例如，实验因子（维度）的影响不能完全分离。此外，由于方差的减小，我们只能用相对较低的精度来估计单一维度的影响。

一般来说，更大的样本能更好地保存全集的特性。但是向调查对象呈现大量的虚拟情境会导致疲劳、无聊和不受欢迎的方法效应，例如启发式回答（Sauer, Auspurg, Hinz, & Liebig, 2011）。一种解决方案是使用不同版本的问卷，把虚拟情境样本划分为不同层（decks）[也称“集合”（sets）或“区组”（blocks）]，呈现给不同的调查对象。例如在先前的例子中，我们可以使用一半的全集（180 个虚拟情境），并把这个样本分为 10 个不同的层，每层含有 18 个虚拟情境。然而，正如第 3 章第 3 节中更详细的说明，随着不同问卷版本（层）数量的增加，每个层的调查对象数量减少，调查对象和虚拟情境特征的混淆风险也在增加。这种混淆因子风险的提高直接危害了对单一维度影响的因果解释。为了避免这种问题，我们可以扩大研究样本，但这需要投入大量的研究资源（如调查时间和金钱）。

总而言之，研究者必须在定义全集大小的维度和水平数量、从全集中抽取的虚拟情境的数量、对不同的虚拟情境层

的分配，以及调查对象样本大小方面做关键决定。所有这些决定都需要在实验的理想特性（例如估计影响实验因子或维度的精度）和进行实验需要的研究资源间进行权衡和取舍。

为了避免误导性结果，我们建议在每个因素调查研究项目开始前都要清晰定义期望测试的研究假设，明确指定期望测试的理论模型（可能是一个具体的回归方程），并使用实验设计的复杂知识来建构虚拟情境样本，以最小的资源实现研究目标的最优化。以下对实践步骤的解释强调了实验设计所涉及的技术。实验设计指的是进行实验的计划，特别是所使用的实验因子（维度）和水平，以及它们的结合。具体而言，术语“实验设计”被用于描述实验部分的特征或研究中使用的虚拟情境。

小结

研究人员必须在实验方法的不同目标间找出一条最佳的途径，例如我们希望估算参数的精度和调查中使用资源（虚拟情境和调查对象的数量）。接下来的章节将指导研究者在这些关键的权衡因素中做出适当的决定，并特别关注实验设计的文献。

第1节 | 选取维度和水平

因素调查研究设计的第一个任务是决定使用哪些变量的维度和水平。因素调查的主要目的是检验社会理论。为此,我们应该首先从经验和理论文献出发,阐明研究假设,并选择适当的情境或对象并以虚拟情境的形式呈现,进而使研究假设和理论概念在研究中得以操作化。[4]在前例研究很少的情况下,我们也可以使用焦点小组选择最重要的维度来评估所关心的问题。因为只有在调查对象必须平衡许多维度的情境中,所设计的虚拟情境才可能揭示调查对象的态度和决策原则。因此必须避免在其他所有维度上占主导地位的维度。

此外,我们必须确定要设计多少不同的维度。在目前的应用中,维度的数量为3个(Berk & Rossi, 1977)到26个(Shlay, 1986)。一方面,过多的维度可能会使调查对象负担过重,并导致无效的实验效果。另一方面,当只向调查对象呈现有限数量的维度和若干虚拟情境时,调查对象就必须反复评估非常相似的虚拟情境,这可能很快就会导致无聊或疲劳,或者虚拟情境可能会错失可靠判断所需的重要信息(Auspurg, Hinz, & Liebig, 2009; Sauer et al., 2011)。

最好的折中方法可能设计大约 7(加减 2)个维度的中等复杂性。方法研究表明,对这样的设计判断具有最大一致性——通过测量在多元回归中可被解释的变异(方差)来判断(Auspurg et al., 2009; Auspurg et al., 2014; Sauer et al., 2011)。虽然对 5 个维度的回应似乎比对 8 个维度的稍有不一致,但在 12 个维度中更不一致。这一发现在年龄越大(≥60 岁),教育水平越低,或不太熟悉因素调查研究的话题的调查对象中尤为明显。因此,我们建议维度的数量不超过 7(加减 2)个,特别是对于年龄越大、受教育水平越低的调查对象来说。这个数量建议也获得了认知科学研究结果的支持。人类在短时记忆中一次最多可以储存大约 7 条信息(G. A. Miller, 1994)。一般来说,要根据研究目的,在需要的情况下才采用更多的维度。当我们只对几个维度感兴趣时,应该采用更小数量的维度。在这种情况下,我们应该对每个调查对象只使用几个虚拟情境(例如只有一个或不超过 5 个虚拟情境)。否则,可能不会有足够的变量来防止无聊和疲劳的影响(对于信息不足的研究,见 Auspurg et al., 2009)。此外,只使用几个变量但每个调查对象都有若干个虚拟情境的情况会增加调查对象对实验操纵的意识风险,这可能会进一步引发不想要的方法效应,例如社会期望偏差(对于这些具体的影响,见第 4 章和第 6 章第 2 节)。

对调查对象而言,获取他们必须评价的对象和情况的清晰印象是至关重要的。在理想情况下,调查对象不需要建构任何丢失的信息,因为这会导致较低的测量信度。验证研究假设所需的维度必须包含在虚拟情境内。当每个调查对象面对多个虚拟情境时,在虚拟情境模块的介绍部分中应该详

细说明所有不需要变化的信息，保持虚拟情境的简短明了。例如，在介绍中，我们可以告诉调查对象在虚拟情境中的所有员工都是全职的，同时增加一些关于水平定义的解释，例如工作任期是“短”或“长”。在维度和水平上的精度是非常重要的；调查对象对文本理解的困惑可能会影响对效应量的估计。

在明确维度之后，我们必须确定每个维度的水平。回想一下，水平的数量定义了虚拟情境全集的大小。一般来说，更有效虚拟情境样本允许通过更小的全集来获得更精确的参数值。因此，尽可能地限制水平的数量是可取的。对于大多数维度来说，仅使用两个水平来检验研究假设就足够了。自始至终使用两个水平可能进一步帮助避免水平数量带来的影响(number-of-levels effects)。研究相关的调查方法——联合分析和选择性实验——以更多水平的维度会吸引调查对象的更多注意，由此对判断的影响也比那些水平更少的维度更大(例如，见 Wittink, Krishnamurthi, & Nutter, 1982)。如果这种水平数量的影响不是由调查对象的态度所证明，那么它就代表了应该避免的方法论的产物。避免这种偏见最简单的方法是在所有维度上采用相似(“平衡的”)的水平数量。另一个平衡水平数量的基本原理是：获得更好的统计学意义上更有效的虚拟情境样本的可能性(见第3章第2节)。此外，当在维度上使用相同的水平数量时，我们可以很容易地比较不同维度间的回归系数(见第5章)。

为了估计维度和结果变量之间的非线性关系，我们至少需要三个水平。在我们的例子中，我们预期在年龄和评

价之间存在非线性关系。年龄大的员工理所当然被视为收入更多，但是这种影响可能随着年龄的增加而减少。如果只使用两个水平的年龄，调查者就不能识别这种类型关系（Fox，2008）。关系越复杂，所需要的水平就越多。例如，“S”形的关系模型至少需要四种水平（以便可以使用回归样条）。

另一个采用两三个以上水平的原因是为了研究更多水平数量和建构更具有现实性的虚拟情境。例如，在收入公平的研究例子中，我们可能对涵盖了全部声望的广泛职业分析感兴趣。然而，在广泛的职业中，最好提供两个以上的收入水平来建构更贴近现实的虚拟情境。基于上述原因，这个例子中我们在职业和收入维度中设计了十个水平，而其他维度则限制在两三个水平以内。

最后，所选择的维度和水平应该使虚拟情境案例具有可读性和可信性。为了获得第一印象并准备随后的虚拟情境文本编写，可以绘制如表 3.1 所示的单个文本短语的框架。在第一稿中，文本通常不太流畅，所拟的虚拟情境案例读起来也不太顺畅。在这种情况下，你可以再确定维度和用词。有时会出现不符合逻辑或不可信的情况。不同维度的某些水平也可能不合逻辑。例如，长时段的工作任期和没有工作经验的组合是不符合逻辑的。不可信的例子指的是那些在现实中很少发生并可能刺激调查对象的例子，例如非同寻常的极高收入或极低收入的职业（例如没有技术含量的工作者每月净收入 10 000 欧元）。当从全集中抽取虚拟情境时，这类组合会被排除在外（见第 3 章第 2 节），但最好提前避免它。

表 3.1 虚拟情境的维度、水平和文本用语草稿

	维度	水平	虚拟情境文本
1	年龄		A
		1	30
		2	40
		3	50
		4	60
			……岁
2	性别	1	男
		2	女
3	教育水平	1	没有学历
		2	受过职业训练
		3	有大学学历
4	职业		工作是:
		1	没有技术的工人
		2	门卫
		•	•
		•	•
		•	•
		10	医生
5	经验		拥有……
		1	很少
		2	很多
			工作经验
6	工作任期		他或她在……
		1	短
		2	长
			时间以前加入公司
7	子女		有……
		1	没有
		2	一个
		3	两个
		4	三个
			孩子
8	收入		他或她的每月净收入是……
		1	500
		2	950
		•	•
		•	•
		•	•
		10	15 000
			税前欧元

小结

根据理论，在理想情况下我们应该选取7（加减2）个变量维度，并为每个维度指定两到三个水平。更多或更少的维度会影响判断的一致性和可靠性，尤其是对年长且受教育程度较低的调查对象，或那些对研究主题不熟悉的人而言。少量的水平产生更有效的虚拟情境样本，而平衡数量的水平有助于避免水平数量影响。需要两个以上的水平才能估计非线性关系。有时为了达到特定的研究目的也需要设计更多的水平。最后，我们应该测试维度和水平是否可以组合成合理的虚拟情境案例。因此，在设计因素调查时，首先要确定虚拟情境包含的最重要的维度，并尝试使用有限的、平衡的水平数量。

第2节 | 实验设计

在本节中，我们假设已经明确了维度和水平，并希望基于这些条件来设计因素调查。如上文所述，在大多数情况下，维度和水平组合的可能数量（虚拟情境总集）太大，以至于无法在一个调查中完成。因此我们通常不得不使用全集中的一部分。到目前为止，因素调查研究大多使用随机抽样（Wallander，2009）。实验设计的文献表明，随机的选择可能会产生估计偏差（伪相关关系引起的忽略变量偏差，例如，混淆交互作用，下面将提供更多细节），内在而言是低效的，因为比起有效的抽样方法，它们会导致更低精度的参数估计。在达到同等精度的情况下，虚拟情境抽样技术更有效，所需要的调查对象或每个调查对象所面对的虚拟情境更少，同时可以更好识别所有增益参数（parameters of interest，例如需要检验研究假设的回归系数）。这些目标标准对于一个好的实验设计至关重要，尤其是对资源最小化（调查对象的数量）感兴趣的人而言。任务就是识别出全集中的部分（维度和水平的组合），从而最大程度地获取信息。

当规划详细的实验设计时，研究者应该首先明确自己想要估计的参数。当只使用所有可能的虚拟情境中的一部分时，一个最优化的设计需要具备所有想要识别的参数，包括

虚拟情境之间所有可能的交互关系、其他虚拟情境变量，甚至是调查对象变量。因素调查方法估计的模型大部分都是可加模型，因为人们期望观察到的结果是单维度的线性组合（以及它们的相互作用，即在参数中是线性模型）。此外，人们假定在判断中存在一些随机变化，这是在结果变量中反映出所未测量出的影响。[5]例如，假定（非）正义性收入的量是职员特征的线性函数。这一假设在回归方程 3.1 中被公式化，X 代表 p 虚拟情境的维度，Z 代表 q 调查对象变量，β 和 γ 是各自的回归系数，i 是单个虚拟情境的系数，j 是单个调查对象的系数，n_d 是每个调查对象面对的虚拟情境的数量，而 n_r 是调查对象的数量。判断中的随机成分用 ε_{ij} 表示：

$$Y_{ij}=\beta_0+\beta_1 X_{ij1}+\cdots+\beta_p X_{ijp}+\cdots+\gamma_1 Z_{j1}+\cdots+\gamma_q Z_{jq}+\varepsilon_{ij}$$

$$\text{其中}, i=1, \cdots, n_d; \ j=1, \cdots, n_r \qquad [3.1]$$

在单一调查对象评估若干个虚拟情境的情况下，可能会在调查对象水平上增加一个额外的误差项来处理多水平数据结构（第 5 章）。当在若干个虚拟情境之间估计选择时，公式可能会不同，因为只有感知刺激（例如，每个虚拟情境的效用水平）可以被假设遵循一个参数是线性的函数，而最好选择逻辑函数或概率函数（Hensher, Rose, & Greene, 2005; Ryan, Gerard, & Amaya-Amaya, 2008；也可以参见第 6 章第 1 节）。然而在大多数情况下，结果变量是一个区间标度变量（interval-scaled variable），因此可以使用线性函数和普通最小二乘法（OLS）回归。

在这个步骤中，重要的是确保考虑了所有用于验证研究假设的变量。必须排除交互作用项，尤其是在虚拟情境维度

间的这些交互作用项吗？我们应该涵括所有的平方项来拟合维度和结果变量之间的非线性关系吗？在虚拟情境维度和调查对象特征之间要考虑什么跨水平的交互作用吗？

下面，我们首先简单演示一下为什么抽样总会导致信息丢失（下文第一部分），同时我们会介绍实验文献的一些术语（第二部分），然后介绍用于编写虚拟情境案例的各种技术，主要关注部分因子（fractional factorial），*D*-efficient 设计（第三部分）。在结语中（第四部分），总结了实验设计最重要的问题。

全因子和虚拟情境部分

为了说明由于部分使用虚拟情境全集而导致信息丢失，我们看一个简化的收入公平的例子。[6]假设我们只对以下三个维度感兴趣，每个维度都有两个水平：女性（是/否）、已婚人士（是/否）、高工作投入（是/否）。我们可以预期这些维度间存在交互作用，例如，婚姻状况对男性和女性员工的影响可能不同，因为性别角色能预测已婚男性要挣钱养家（“男性是担负生计的人”）。相较于其他人而言，较低的工作投入对已婚女性的负面影响比其他人低，因为已婚女性可能有孩子要照顾，并且要承担更多的家务，这是她们工作投入较低的一个正当理由。在以下回归方程中，正义的评价 Y 与虚拟情境维度之间的关系假设如下所示：

$$
\begin{aligned}
Y = \beta_0 &+ \beta_F \text{女性} + \beta_M \text{已婚} + \beta_E \text{工作投入} \\
&+ \beta_{FM} \text{女性} \cdot \text{已婚} + \beta_{FE} \text{女性} \cdot \text{工作投入} \\
&+ \beta_{ME} \text{已婚} \cdot \text{工作投入} \\
&+ \beta_{FME} \text{女性} \cdot \text{已婚} \cdot \text{工作投入} + \varepsilon
\end{aligned}
\qquad [3.2]
$$

注意:这个方程包含了三个双向交互项(在第二行)和一个三方交互(女性·已婚·工作投入)。

这个虚拟情境全集是由三维度两水平构成的,其大小是 $2^3=8$。表 3.2 展示了完整的阶乘。维度被重新命名为 D_1、D_2、D_3,两个水平被编码为 -1 和 $+1$("正交"编码)。此外,所有可能的交互作用都展示出来。通过正交编码,将维度的水平相乘即可得到水平(详见 Kuhfeld, 1997)。例如,D_1、D_2 的两个较低水平(-1)相乘得 $+1$,交互项的水平为 $D_1 \cdot D_2$。

表 3.2 2^3 全因子和两个分部

	主效应			两向交互			三方交互	
序号	D_1	D_2	D_3	$D_1 \cdot D_2$	$D_1 \cdot D_3$	$D_2 \cdot D_3$	$D_1 \cdot D_2 \cdot D_3$	
1	−1	−1	+1	+1	−1	−1	+1	A分部
2	−1	+1	−1	−1	+1	−1	+1	
3	+1	−1	−1	−1	−1	+1	+1	
4	+1	+1	+1	+1	+1	+1	+1	
5	−1	−1	−1	+1	+1	+1	−1	B分部
6	−1	+1	+1	−1	−1	+1	−1	
7	+1	−1	+1	−1	+1	−1	−1	
8	+1	+1	−1	+1	−1	−1	−1	

注:该例子改编自 R.F. Johnson et al.(2006)。

完整阶乘的矩阵是正交的,表示单独的列之间不相关。这是一个理想的设计矩阵,它使研究者可以估计单个维度的独立影响。矩阵显示了水平的平衡,表明了所有水平都以相同频率发生(水平 -1 和 $+1$ 都以 50%的概率发生)。对于实验设计来说,这是一个理想的特征,因为它表明存在可以估

计单个维度的影响最大方差。这个特征反过来带来了最低标准误，进而保证了最大水平的参数估计的精度。此外，完整因子允许识别所有主效应和交互作用项。

如果只采用虚拟情境全集的一部分会出现什么情况呢？在我们的例子中，虚拟情境全集是相当小的。但是如前所述，大多数实际的应用包括更多的维度和水平，因此产生了全因子。在实验文献中，任何处理方式的子集叫部分因子设计。在表 3.2 中，我们可能只使用组合维度的前半部分（在表 3.2 中的 A 分部）。在这个部分中 D_1 水平与 $D_2 \cdot D_3$ 水平完全相同。统计学上，这两个回归量是完全相关的，它们的影响是不可分割的。在实验术语中，这两个回归量是混淆的或别名的，即 $D_2 \cdot D_3$ 是 D_1 的别名。这同样适用于 D_2 和交互作用项 $D_1 \cdot D_3$，以及 D_3 和 $D_1 \cdot D_2$。此外，截距 β_0 完全与三方交互 $D_1 \cdot D_2 \cdot D_3$ 混淆。当采用这部分时，以下的两个回归方程是等价的（比较 R.F. Johnson et al.，2006：162）：

$$Y=\beta_0+\beta_1 D_1 \qquad +\beta_2 D_2 \qquad +\beta_3 D_3 \qquad +\varepsilon \quad [3.3a]$$

$$Y=\beta_0+\beta_1 D_2 \cdot D_3+\beta_2 D_1 \cdot D_3+\beta_3 D_1 \cdot D_2+\varepsilon \quad [3.3b]$$

所有主效应的系数只有当我们假设混淆的双向交互作用是可以忽略（有零效应）时才能够辨识出来，反之亦然。在没有关于这个假设是否成立的信息下估计方程 3.3a，就会导致估计的偏差，因为回归系数不能单独测量主效应的影响，它们还测量了别名的交互作用的影响。

在实验术语中，A 分部和 B 分部都表示正交主效应的设计。这样的设计在假设所有交互作用是零的情况下只允许

对主效应进行评估。它们被频繁地运用于联合分析和市场营销研究。对于因素调查在社会科学中的大部分应用，我们有充分的理由假设至少有一些双向交互作用的应用。之前描述过一个三方交互作用的例子（女性·已婚·工作投入）。使用主效应设计，这些交互作用的假设不能够被验证。甚至对主效应的估计是有偏差的，这种偏差是由于遗漏条件未被发现的影响的伪相关所导致的，这会带来极具误导性的结论。

在此期间，从实验文献中引入以下术语是非常有帮助的：设计分辨度（design resolution）。这个分辨度定义了哪些效应可以通过使用虚拟情境部分识别出来，包括可能存在的交互作用（Dülmer，2007；Kuhfeld，1997）。在分辨度Ⅲ的设计中，只有主效应是相互独立进行估计的。在分辨度Ⅳ中，研究者提前指定的一些双向交互作用可以被独立估计，而其他的交互作用都是相互联系的。在分辨度Ⅴ中，所有主效应和双向交互作用是可识别的。[7]更高层次的分辨度通常需要更多数量的虚拟情境。对于许多社会科学领域中的研究而言，我们期望至少使用分辨度Ⅴ的设计，因为我们必须预测双向的交互作用。在研究领域中，关于可能进行交互作用的潜在信息是可获得的，我们可以明确一些双向的交互作用而忽视其他（分辨度Ⅳ的设计）。但是，我们必须意识到，如果我们不能相信这些混淆的交互作用是可以忽略的，那么这种方法可能会导致有偏差的结果。

鉴于上述考虑，应该确立以下声明：使用虚拟情境的一部分而不是全集的代价是常常丢失信息。除了完全混淆之外，抽样会导致部分混淆（即维度间适中的或很强的相关；

Atzmüller & Steiner, 2010)。因此,我们应该放弃那些确定对结果没有任何影响的高阶交互作用。三方和高阶交互作用是罕见的(Louviere, 1988);因此它们可以被忽略。但是在社会科学中经常发生的双向交互作用应该识别出来。[8]

有效设计

设计效率是衡量实验设计的一个指标。在设计实验时,目标是基于给定的适当的调查对象和每个对象面对的虚拟情境基础,选择一个可以提供最多统计信息的方法。也就是我们要寻求关于检验研究假设每个参数的信息最大化。从统计意义上,这等同于使得数据分析(如回归)估计出来的系数的方差和协方差最小化。最小化参数的方差会导致信息的最大化。

统计信息的常用测量是 Fisher 信息矩阵(Fisher information matrix, FIM)。当使用矩阵代数时,这个矩阵可以被记为 $\mathbf{X'X}$, $\mathbf{X}$ 表示虚拟情境变量的向量(Kuhfeld, Tobias, & Garrant, 1994)。设计效率的几个衡量指标都依赖于 FIM,最显著的是 *D*-efficiency。*D*-efficiency 同时反应了正交性和水平平衡性(即所有水平的相同频率)。这两个标准的优化有助于提高统计分析中参数估计的精度。n_s 表示在样本或部分中虚拟情境的数量,$|\mathbf{X'X}|$ 表示虚拟情境变量(包括截距)中 FIM 的行列式,而 p 表示与回归系数的数量(包括截距)相关的维度和维度之间的交互作用,这是我们希望识别的。*D*-efficient 的公式如下所示(Dülmer, 2007; Kuhfeld et al., 1994):

$$D\text{-efficiency} = 100 \cdot \frac{1}{n_s \cdot |(\mathbf{X}' \cdot \mathbf{X})^{-1}|^{\frac{1}{p}}} = 100 \cdot \left(\frac{1}{n_s} \cdot |\mathbf{X}' \cdot \mathbf{X}|^{\frac{1}{p}}\right) \quad [3.4]$$

对于给定数量的虚拟情境,较少的维度以这样的方式建构样本,即这些维度之间的相关性不强。同样地,更大数量的虚拟情境的样本(同时定义了参数估计的自由度)有助于减少维度间的协方差(因此增加了相关),从而增加了参数估计效益的精度(如回归系数测量单一维度的影响)。在公式中,FIM 的增加反映了参数估计信息量的增加。当估计 OLS 回归时,回归系数的方差—协方差矩阵与 D-efficiency 成反比。[9]因此,倘若调查对象产生了有效的判断,那么最大化 D-efficiency 相当于在 OSL 回归中获得最精确的参数估计。

当所有维度的编码都是标准化的正交(详见 Kuhfeld,1997),D-efficiency 有一个[0;100]的范围,100 代表最有效的设计(Dülmer,2007;Kuhfeld et al.,1994)。值为 100 时只可能是完全正交的设计,且是所有水平相等的频率。当在不同维度中使用不同水平的数量或是当搜寻小的分部时,通常无法达到这个值。因此,我们不应该使用 D-efficiency 值作为效率的绝对测量标准,而应该将这些值解释为处于考虑中的不同设计的效率的比较值(Kuhfeld et al.,1994)。也就是说,更合理的解释是一个 A 设计的效率是另一个 B 设计的 x%,x 是两者 D-efficiency 的比率(A 的 D-efficiency 除以 B 的 D-efficiency),而不是检查单个设计的绝对值。[10]尽管如此,一些关于如何对 D-efficiency 值进行分类的信息可能会有所帮助。第一个经验法则是识别 90 以上的 D-efficiency

值。在我们的经验中，这样的设计提供了足够的统计效力来满足社会科学的大部分研究目标（在这里，目标很难达到准确的预测参数值，而是有足够的统计检验力验证社会理论）。如果排除了不合情理或不合逻辑的例子（第 3 章第 4 节），人们可能会对较低的 *D*-efficiency 值感到满意。我们在第 3 章第 5 节提供了基于模拟的说明。

正如前文指出的，更有效率的设计意味着用更少的虚拟情境（即调查对象、每个调查对象面对的虚拟情境，或者二者的结合）就能取得与较低效率设计一样的统计检验效力。一个经验法则是，A 设计的设计效率是 B 设计的 80％，我们需要多于 25％的（相当于 80％的倒数，即 1/0.8＝1.25）虚拟情境判断来达到与更有效率的 B 设计（Chrzan & Orme, 2000）一样的参数估计的精度（标准误大小）。无论如何，即使可以负担更大的样本量，我们还是青睐更有效率的设计，因为它们提供了更多的空间来测试不同的实验条件（比如，进一步的维度或交互作用；Scarpa & Rose, 2008）。因此，我们推荐能识别所有相关参数估计并提供最大的 *D*-efficiency 的设计。

构建有效率设计的方法

正如之前讨论的，好的实验设计标准如下：设计应包含足够识别所有增益参数，同时提供最大的统计效率的信息。正交且平衡的设计提供了最大效率（从单维度水平以相同频率发生）。研究者应该试图同时优化这些标准。

建立计划

你可以参考教科书或网站发布的设计题录建立计划(见 Atzmüller & Steiner, 2010; Gunst & Mason, 1991)。图 3.1 提供了两个建立计划的例子。在使用这些计划时,必须将数字代码转化为虚拟情境维度的表述(见第 4 章和第 3 章第 1 节的表 3.1)。

下面的计划展示了如何将三个有着两个水平的维度(2^3 设计)与一个带有两个和一个带有三个水平的维度($2^1 3^1$ 设计)结合起来,以获得正交的主效应设计(分辨度Ⅲ)。对于 2^3 设计,一个子集只包含全集($n_v=4$)中虚拟情境的一半,足以达到正交的主效应的设计(见左边的计划;在这个子集里,所有三个维度间是不相关的)。与此相反,在 $2^1 3^1$ 的设计中,完美的正交性(维度之间零相关)在使用整个因子("全集";见右边的计划)时可以被保存。从技术上讲,对于右边的 $2^1 3^1$ 设计,没有正交因子存在(即不可能保留与任何全集的子集不相关的维度)。这两种设计不仅是完美的正交,而且是平衡的(所有水平以相同频率出现)。这是理想的设计特征,因为这个设计保证了维度水平方差的最大化,以及对虚拟情境影响估计的精度。

2^3				$2^1 3^1$	
1	1	1		1	1
2	1	2		1	2
1	2	2		1	3
2	2	1		2	1
				2	2
				2	3

注:这些例子改编自 Kuhfeld(2010)。

图 3.1 设计计划的案例

然而大多数题录只提供正交的主效应设计(分辨度Ⅲ)。这些题录主要用于市场营销研究(通常是联合效度)。与因素调查在社会科学中的应用相比,市场营销研究认为交互作用影响轻微,所以常常忽略了它。许多计划仅限于两三个水

平的维度。一些技术涉及如何修改这些计划以在设计分辨度或使用水平的数量方面达到更大的灵活度[例如“折叠方法”(foldover approach); Gunst & Mason, 1991]。然而,这些技术和标准题录并不必然提供最有效率的设计。它们通常只优化正交,不考虑水平平衡。此外,它们在关于所使用的水平和维度的数量方面表现出较低的灵活性(R. F. Johnson et al., 2006)。设计题录中没有提供避免不合理或不合逻辑的情况的建议(关于这些情况的更多信息在第3章第4节中提供)。

计算机算法

与设计题录相比,计算机算法提供了更大的灵活性,并提供了更有效率的设计。计算机算法评估了数千甚至数百万个感兴趣的设计方案,然后为给定的统计效率标准选择最有效的设计(Kuhfeld, 1997; Kuhfeld et al., 1994)。

算法运用先要明确希望识别的全因子(维度和水平的数量)和参数(主效应和交互作用)。此外,必须指明我们期望使用的 n_s 部分(虚拟情境的数量)。这些虚拟情境的数量代表了我们计划运行的不同实验的数量。在实验文献中,不同刺激经常被称为测试(runs),因此,样本中虚拟情境的数量(我们用 n_s 表示)与测试的数量相类似。这个数量必须超过参数的数量(维度和交互作用),有一个自由度的估计(关于样本大小的更多信息,见第3章第5节)。这个算法首先根据这些明确条件(可以从题录中设计)建立一个“候选集”,然后搜索更高 *D*-efficiency 分数的设计。[11] 有着最高(或至少是足够高)的 *D*-efficiency 分数的设计会被作为数据集。

免费软件宏“*%Mktex*”(发音为“maktex”)为 *D*-efficient

设计提供了著名的复杂算法集。这个宏是由沃伦·库菲尔德(Warren Kuhfeld, 2010)在软件包SAS中编写的。"*%Mktex*"的宏采用了若干种不同的搜索算法,考虑了大量的设计题录,而且提供了极大的灵活性。例如,我们可以定义在探寻效度设计的过程中,达到的最小效率值或是应该投入的最大时间。程序允许对应该避免的水平组合具体化(更多关于不合逻辑的组合在第3章第4节讨论),同时它还可以与其他宏结合起来,例如生成理想的样本大小(虚拟情境的数量)。SAS宏的另一个优点是,它们可以显示混淆参数。当使用一个特定部分时,我们可以准确知道丢失的信息。所有这些宏都在SAS的官网上提供,同时还有全面的文档。[12]其他软件工具,例如免费软件R的用户编写包也是可获得的。更多关于可用软件分辨度的信息以及"*%Mktex*"语法代码的例子可以在以下网站上找到:www.sagepub.com/auspurg_hinz。

接下来,我们讨论一下在使用计算机搜索时,如何提高*D*-efficiency(更全面的内容见Kuhfeld, 2010)。通常,单维度的水平数量多少有点任意性,因此通过使用有些不同的规格,因素调查实验的目标仍能实现。所以我们可能会验证例如不同数量的水平、维度或交互作用,以确定这些变化如何影响效率。如果一个小的修改导致了效率的极大提高,那么我们可能会考虑使用这些不同的规格组合。当维度与水平数量相等或当水平数是彼此的倍数时,设计是有效率的。例如,与$3^2 2^3 5^1$的设计相比,我们会使用$3^2 2^3 6^1$的设计。即使后者会导致一个更大的虚拟情境总集,但可能得到更多有效的分部。此外,我们可以测试不同虚拟情境分部的大小,并看看这些变化如何影响效率。在大多数情况下,一项研究使

用多少虚拟情境是具有一定灵活性。我们可能通过试错或者查找设计文献中如何最大化效率的进一步建议来测试这些修订(例如 Gunst & Mason, 1991; Kuhfeld, 2010)。如果我们使用了一些小的分部并探索了极高的设计分辨度,那么通过这种修正得到的 *D*-efficiency 的改进就很可行了。

对于许多非标准化设计(设计中包括大量的维度水平、复杂的交互或避免不合情理的案例这样的限制),计算机算法为实验设计生成提供了唯一可行方法。正如前面提到的,修改设计题录的技术是存在的,但非常耗时并容易出错。出于这些原因,我们强烈推荐使用计算机算法。

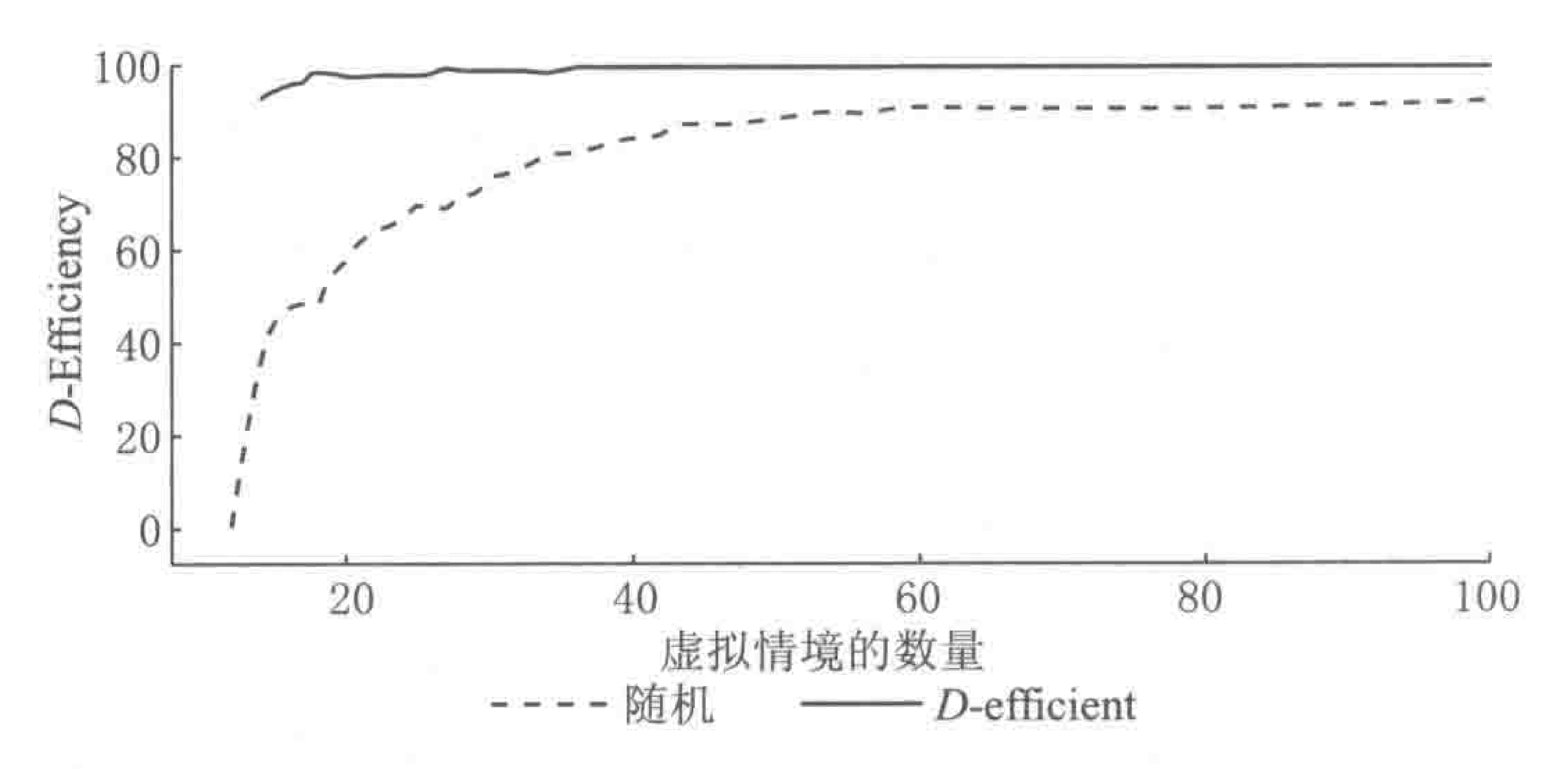
随机样本

迄今为止,大多数因素调查研究不依赖于 *D*-efficient 分部,而是采用对虚拟情境的随机样本(Wallander, 2009)。这里有多种生成随机样本的方法,例如对有或无替代的单一调查对象提供虚拟情境(没有替代的调查对象更有效,因为这样会使用更多不同的虚拟情境)。然而,所有这些技术实际上与临时组合维度一样有效。

只有一个统计学上的理由来解释为什么青睐随机样本而不是分部样本(Gunst & Mason, 1991)。有人可能会说当使用随机样本时,所使用的水平和它们的结合方式可以被解释为能代表所有水平组合的更大虚拟情境总集,从而使研究者得出更普适性的结论(Alve & Rossi, 1978)。然而这种说法并不令人信服。我们怎么知道那些除随机采用的值以外的值域对虚拟情境判断有同样的(线性的)影响呢?毕竟,对无法观测的数据点或值域进行推论总是存在问题的。[13]

实证研究已经证明，就设计效率而言，分部样本优于随机样本(R.F. Johnson et al., 2006; Kuhfeld et al., 1994)。[14]这一发现是因为 *D*-efficient 样本是被用来提高设计效率的，因此需要随机抽取一个样本，使它优于 *D*-efficient 样本。

使用随机样本的另一个原因是，它们比 *D*-efficient 样本更容易实现，因为它们不需要专业知识和软件来进行实验设计。统计理论和模拟研究都表明，通过增加样本量，分层和随机抽样之间的差异会减小(Dülmer, 2007)。这种现象在图 3.2 中显示，在一个有着七个维度、两或三个水平(一个 $3^3 2^4$ 设计)的典型的虚拟情境设计中，*D*-efficient 和随机样本的 *D*-efficiency 值被显示出来。样本量(n_s)从左到右增加(从 15 到 100 个虚拟情境)。随着样本量增大，两个抽样方法的 *D*-efficiency 值越来越接近。



注：对应每个虚拟情境分部的大小，各提供一个随机样本和 *D*-efficient 样本。感谢 Carsten Sauer 为 *D*-efficient 样本提供 Stata 的编程代码。

图 3.2　不同虚拟情境分部大小的随机样本与 *D*-efficient 样本的 *D*-efficencies 值($3^3 2^4$ 设计的正交主效应)

尽管我们的例子中(没有任何正交的交互条件)只用正交的主效应(分辨度Ⅲ)设计，当随机样本达到 100 个虚拟情

境时只达到了 D-efficiency 最大值的 92.8（见图 3.2）。相反，采用分部 D-efficient 样本时，可能达到的最大效率值（$D=100$）在 $n_s=36$ 虚拟情境时就已经出现了（基于 D-efficient 和随机抽样技术的典型设计的 D-efficiency 分数将在第 3 章第 5 节呈现）。

随机样本和分部样本之间的效率差异取决于几个设计特征，例如维度和水平的数量，以及识别增益的交互项数量。因此，为了确定不同的分部样本在统计效率方面优于随机样本的程度，我们需要进行模拟研究。在更大的样本中，随机样本的低效率问题消失了。在许多实际应用中，一个大约 200 个虚拟情境的样本足以通过随机抽样技术获得有效的样本（例子见第 3 章第 5 节的表 3.3）。然而，我们必须记住，随机样本总有其代价的。从定义上看，随机样本导致了偶然发生的参数混淆是。人们通常不知道哪些维度是混淆的，哪些信息丢失了。这个缺陷在社会科学的应用中尤为严重，因为人类的决定、判断或定义的研究往往过于复杂，以至于无法预先排除任何双向交互作用。因此对于较小的分部（例如，在有 7 个维度少于 200 个虚拟情境的标准应用中），我们不推荐使用随机抽样技术。如果使用，就应该仔细检查它们的效率（正交性和水平平衡），同时评估是否存在任何可能妨碍参数识别和导致偏差估计（由于带有混淆维度或交互作用的未定义的伪相关而忽略的变量偏差）的混淆。[15]

总而言之，对增益参数的识别不应该完全留给偶然性（Carson, Louviere, & Wasi, 2009）。至少当使用带有高分辨度的小分部或复杂设计时，我们强烈推荐使用分部 D-efficient 设计。

附记:类别结果的调试

读者如果有计划使用标准回答量表可以跳过这部分。如果你计划使用分类结果、运用逻辑回归或概率回归进行数据分析,以及对高水平效率(参数估计的精确度)感兴趣,我们可能会为非线性(回归)模型研究几个专门的 *D*-efficient 设计。

注意计算 *D*-efficiency 的公式 3.4(见第 3 章第 2 节)是基于该模型的参数估计是线性的假设,正如使用 OLS 回归一样(Scarpa & Rose, 2008)。在参数模型不是线性的时候,例如逻辑或概率模型,用来估计类别结果的(例如,在虚拟情境或回答类别描述的不同行为选项中进行选择)最佳实验设计就变得更加复杂。[16]原因是从 FIM 的应用到非线性模型是不同的公式,FIM 现在是真实回归系数矢量的另一个函数(Ferrini & Scarpa, 2007; R.F. Johnson et al., 2006)。使用更简单的线性模型公式等同于假设所有的参数估计是零。然而这是一个非常不现实的假设。因为所选择的因素调查维度已经被研究者假设为对虚拟情境估计有很大影响的因素。由于因素调查主要选择线性模型(这是该方法和其他实验调查方法之间最重要的不同之一),所以我们只提供了一些如何对非线性模型进行有效设计的初步建议。感兴趣的读者可以详见第 6 章第 1 节,找到这些方法(包括文献推荐)的额外建议。此处只提出与实验设计相关的基本信息。

如前所述,非线性模型的 *D*-efficiency 取决于真实系数的未知参数向量。因此,将非线性模型效率最大化的主要方法是使用预先确定的参数[被称为先验(priors)],这可能源于现有的研究、理论考虑或可能是最可靠的资源的预测试项

目。我们推荐这些包括先验在内的多种策略，包括“贝叶斯设计”，允许研究者在他们参数值的先验知识中解释一些不确定的内容(Ferrini & Scarpa, 2007)。有兴趣的读者可以参考这些非线性模型设计的文献(Ferrini & Scarpa, 2007; Scarpa & Rose, 2008)。幸运的是，许多软件包(例如用于选择实验的 SAS 宏和 Ngene 软件)允许在算法中使用这些先验。

关于非线性模型的这些设计策略的研究仍在进行中。一些最初的模拟显示，使用贝叶斯 *D*-efficient 设计或基于先验的其他设计可以提高非线性模型的设计效率，但效率值的提高常常是微不足道的(Ferrini & Scarpa, 2007)。因此，许多研究者认为，在缺乏高质量先验信息的情况下，可以为线性模型选择更简单的 *D*-efficiency 测量(Ferrini & Scarpa, 2007)。[17] 大多数因素调查被应用于验证感兴趣的理论，而不在于实现参数估计的最大精度(在消费者研究中则试图进行尽可能精确的预测)，线性模型标准公式的 *D*-efficiency 应该足够了。

关于实验设计的结语

“研究者应该认识到所选择的设计……至少与那些人们用来分析结果数据的模型同样重要。”(Louviere, 2006:177)无论使用哪种实验设计，我们应该在实施之前仔细验证和评估设计的特性，以确保设计能够达成所有的研究目标(Amaya-Amaya, Gerard, & Ryan, 2008)。在收集数据之后就没有办法改进研究设计了。尤其是如果我们使用大的全

集中的小分部,实验设计必须得到充分的重视。有时软件包会直接提供混淆参数以及与设计质量相关的其他标准的信息。如若不然,可以通过检查所有维度和交互作用的相关矩阵(双向的和可能更高阶的交互作用)来获得初步理解。请注意,一些混淆可能无法通过双变量相关性检测到,可能存在三个或更多维度的线性组合产生的多重共线性。

我们应该总是不断提醒读者所使用的实验设计类型(Amaya-Amaya et al., 2008)。读者必须知道统计的特性和虚拟情境样本的局限,尤其是混淆模式(例如设计分辨度和正交的交互条件)。通过标准设计,勾勒出设计分辨度为确定混淆模式提供足够的信息。尽管如此,*D*-efficiency 值信息是有价值的,因为它帮助读者评估设计的统计效力。这些信息至少和回归方程估计时一样有用。关于如何实现 *D*-efficient 样本的关于收入公平的可参考案例(例如通用软件工具的编程代码)能在我们的网页上找到。

小结

实验设计决定了研究的重要方面,如参数估计的可识别性和统计效率。一旦研究者收集了虚拟情境判断,就不能纠正不恰当的样本了。使用虚拟情境总集的分部样本(全因子的分部)会增加一些参数间混淆的风险(主效应或交互作用;对于两个水平以上的绝对维度,这至少对多项影响来说是正确的,它允许识别非线性关系)。只有假设的混淆参数的影响可以忽视时,才可能对参数进行有效解释。我们不建议使用随机样本,因为这为完全混淆模式留下了机会,会限制实

验设计的主要优势——实验操纵影响的明确解释。最优的设计允许识别所有需要验证实验假设并能提供最大统计效率的参数(即所有参数可以最大精度地被估计)。

一些满足目标标准的设计可以在实验设计的题录中找到。但是,在大多数社会科学应用中,最好使用计算机搜索*D*-efficient 设计。对于大多数因素调查应用而言,我们使用那些所有或者部分双向交互能被分别估计的设计(即分辨度Ⅳ或更高层次)。

第 3 节 | 划分区组

研究设计可以被分为不同的层块或区组(也称集合),呈现给不同的调查对象。一个虚拟情境设计有 n_s 个虚拟情境,被划分为 d 等大小的区组,每个区组都包含 $n_d=n_s/d$ 个虚拟情境。因此,不同组的调查对象被要求评估不同的区组(Atzmüller & Steiner, 2010)。这个方法允许更大数量的虚拟情境和更有效率的设计。各种技术(分层设计和随机样本)可以把实验设计分为不同区组。

到目前为止,在因素调查研究中最常用的技术是随机分区块(Wallander, 2009)。然而随机层块会导致标准化和虚拟情境层块统计效率的水平降低。由于案例数量较少,混淆虚拟情境块中的参数比整个虚拟情境样本更多,因此一些效果只能在层块间(而不是在其中)被估计。换言之,一些参数不可避免地会与层块效应(例如,可能会产生于特定的虚拟情境层块,如包含极端案例的层块)相混淆,同时会与评价层块的调查对象的群体特征相混淆。为了使虚拟情境适应层块的统计信息的损失最小化,应该使分区组的设计仅有很少重要参数,比如高阶交互作用这样的设计成为混淆。使用适当的分区组技术,只有 $d-1$ 参数会与单独层块相混淆(d 再次显示了不同区组的数量,见 Atzmüller & Steiner, 2010)。

当所有区组被若干个调查对象评估时，它们不太可能完全与一些调查对象的特征混淆，这些参数在调查对象的估计中仍然可识别。

如果想要估计调查对象特定的参数，例如回归系数或随机斜率，那么对虚拟情境的有效分区组就显得尤为重要。由于对这些估计的案例数量较少(呈现每个层块的虚拟情境数量)，调查对象特定的估计通常显示出较高的可变性，即它们有很大的标准差[对于这种“跳跃 β 问题”(bouncing beta problem)，见 Hox et al., 1991，以及第 5 章]。这种高可变性一部分是由于不同虚拟情境层块的抽样误差引起的。标准化虚拟情境层块的优点是减少这种误差来源，即排除一些由于方法论上的原因(分得不同的层块)在调查对象之间产生的可变性。此外，当单维度之间有较高的相关性，调查对象可能难以区分这些维度，或是意识到这些系数变化不相关。因此，标准化的层块有助于正确区分出调查对象间的随机差异(Dülmer, 2007)。出于类似的原因和为了避免水平数量的影响，在区组中的高度水平平衡也有助于虚拟情境估计中获得的信息最大化。

在某些情况下，个体随机层块被用于所有调查对象。在这种情况下，每个调查对象会收到自己的虚拟情境层块，这样要有更大比例的虚拟情境全集可以被采用(Rossi, 1979; Rossi & Anderson, 1982)。对于很大的虚拟情境全集而言，混淆的问题仍然存在(Atzmüller & Steiner, 2010)。通常来说，随着虚拟情境全集数量 n_s 的增加，可以被识别的参数的数量也在增加。然而，对于给定数量的调查对象来说，更大数量的虚拟情境全部被转化成对单个层块进行评价的调查

对象的数量会更少。因此，区组效应可能被调查对象特征混淆的风险增加了。同样地，有效的分区组技术提高了正确识别调查对象间差异的可能性（包括跨水平的交互作用）。我们在第 3 章第 5 节的样本量部分中会更详细地讨论这些问题。

总之，我们强烈建议使用慎重的分区组技术，而非随机分区组技术，尤其是当只采用几个不同的层块时（例如，只使用 10 个带有 10 个虚拟情境的不同层块）。为了分层块，我们应该使用与虚拟情境抽样（最重要的参数和高统计效应的可识别性）相同的目标标准。事实上，这个目标可以通过使用前面介绍的虚拟情境抽样的计算机算法实现。[18] 我们在网站上提供了一个说明示例（www.sagepub.com/auspurg_hinz）。

小结

实验设计可以被划分为不同层块以呈现给不同的调查对象。这个方法允许使用更大数量的虚拟情境，通常有助于提高统计效率（*D*-efficiency）。对于不同虚拟情境层块的分组应该以一种参数识别和 *D*-efficiency 为导向的方式进行。如果你对识别研究对象间的差异感兴趣，例如回归估计中随机斜率的识别，这些就尤为重要（详见第 5 章）。

第 4 节 | 不合情理且不合逻辑的虚拟情境案例

一些研究者认为能引入不合情理、超越现实的案例（例如，职业收入过高或过低）是因素调查方法的优势（例如，见 Rossi & Anderson, 1982）。调查对象不一定会在某些不合情理的虚拟情境（关于这个问题的讨论，见 Faia, 1980; Rossi & Alves, 1980），但方法论的研究表明，不合情理甚至不合逻辑的案例（例如，没有大学学位的医生）会对数据的效度产生副作用（Auspurg et al., 2009）。在这样的研究中，调查对象可能会较少关注不合情理的维度，事实上，人们可能会说他们不再像之前那样认真对待这些维度。更糟糕的是，调查对象可能会怀疑整个调查的严谨性。因此，我们建议避免出现维度水平上不合情理和不合逻辑的虚拟情境组合。

尤其是在设计有效率的实验时，提前考虑这些限制是非常重要的。有效的设计很大程度上依赖于对设计空间各个角落的观察。虽然从随机设计中移除了这些极端不合逻辑的案例通常会增加变量间的相关性，但带有分部因子的设计使得其他维度也变得紧密相关。此外，我们若扭曲了水平平衡就可能带来有问题的混淆情况。为了保证分部的 *D*-efficient 设计的质量，特别重要的是保证样本中的所有而不

是分部虚拟情境是由调查对象评估的(另见第3章第3节)。

使用计算机算法可以尽可能地将不现实的虚拟情境和其他不满意的维度从候选的集合中排除,以实现有效的设计(Kuhfeld, 1997);我们强烈推荐这种方法。[19]在这些限制条件下,计算机算法所获得的设计效率通常非常好,至少在限制适中时,它代表了给定限制条件下的最高效率(Chrzan & Orme, 2000; Kuhfeld et al., 1994)。

然而,即使是最精心计划的设计限制也必然会导致统计效率的降低。[20]效率的损失取决于具体的设计和被排除的组合数量。总而言之,效率的损失越高,从全集中抽取的分部越小,必须被估计的参数(交互作用影响)的数量就越多。在这个地方尤其有趣的是排除了更多不满意的维度组合。某些情况下,一些限制可能会导致效率的明显损失;在最坏的情况下,计算机算法可能不再汇集出任何有效的样本。因此,我们建议在明确维度和它们的水平时,应当避免出现许多不合情理或不合逻辑的组合。在收入公平的例子中,我们可能会使用适用于所有维度职业范围的收入水平。然而,在许多情况下,研究目的无法避免这种不合情理的组合。例如,你可能对任期是否不受工作经验影响感兴趣,没有工作经验和长任期工作的组合就不是理性的。再者,研究者必须权衡不同的目标,如采用更广泛的不同维度和水平,这可能导致不合情理的组合出现,以及当这些组合必须被排除时要考虑对设计效率的影响。然而以下的建议是很清楚的:应该从候选集合中排除不合情理和不合逻辑的虚拟情境案例以进行有效设计,而不是在(*D*-)efficient 设计被建立(在我们网站上提供的一个例子)之后删除它们。

对于题录设计，禁止某些组合是非常困难的。就这个问题而言，计算机算法有明显优势，*D*-efficient 设计具有高度灵活性（Kuhfeld et al.，1994）。然而，我们应该注意到，所有的设计由于只允许一些（现实的）组合而付出了代价（就效度的损失而言）。

小结

不合情理或不合逻辑的维度水平对数据质量是有负面影响的。当具体化维度和水平，或应该将它们从候选集合中消除以进行有效设计时，应该试图避免出现许多不合逻辑或不合情理的虚拟情境案例。从样本中删除这些案例将影响设计效度和参数的可识别性。

第 5 节 样本量

我们必须决定要使用多少虚拟情境和不同的层块，包括每个调查对象要估计多少虚拟情境。因为数据的收集的代价是不菲的，每个调查对象通常要评估一个以上的虚拟情境。然而在设计的不同分部使用的案例数量间仍存在一些权衡，如图 3.3 所示。

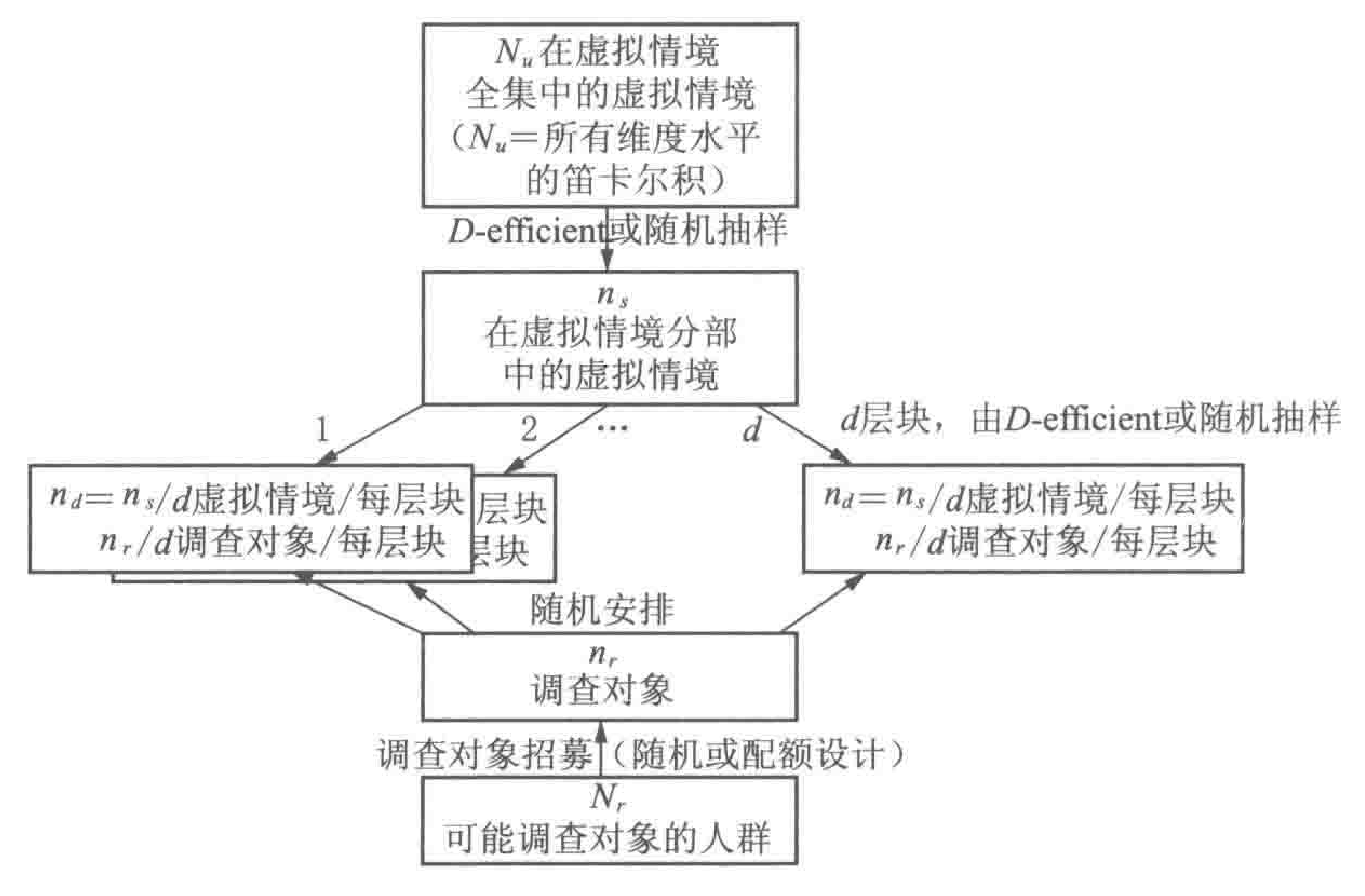


图 3.3　虚拟情境、层块和调查对象的样本容量

随着虚拟情境样本的数量增加（即更高的 n_s 值），找到一个高设计分辨度的高 D-efficient 分部的可能性也在增加。

但是随着样本量的增加，在每个单独层块(n_d)虚拟情境数量确定的情况下，不同层块的数量 d 也在增加；在给定调查对象数量(n_r)的情况下，则分配在同一个层块中的调查对象的数量在减少。下一部分的内容将有助于在这些权衡中确定一个理想的平衡。再接下来的部分提供了关于有意义的样本量的更复杂的统计细节，这是由统计检验力计算(statistical power calculations)来确定的。

样本量的一般考虑

在考虑一项研究中使用虚拟情境数量时，我们首先要考虑的是可以达到的 D-efficiency 的量。高效设计所需要的虚拟情境的数量总是取决于具体规格的应用(维度和水平的数量、避免不合情理的情况，以及因子数量的正交化)。最小化的虚拟情境数量是希望估计参数的数量加 1。然而这种方法会导致一个各维度间高相关性低效分部。在我们的实验中，对于大约 7 个维度的应用，需要大约 100 或 200 个虚拟情境的样本容量(n_s)来达到足够的 D-efficiency。这个概念在表 3.3 中显示了以下四个典型设计的模拟结果：7 个维度的设计($3^3 2^4$ 设计)和 9 个维度的设计($3^4 2^5$ 设计)，二者都允许所有的情况，并排除了一些不合情理的情况。一个不合情理的情况的例子是第一维度的第三水平(X_1，三个水平维度)和第二维度的第二或第三水平(X_2，三个水平)的同时出现，这是仅有的一个适度的限制(技术上而言，不合情理的事例是 $X_1=3$ 且 $X_2 \geqslant 2$)。这里用样本量为 50、100 和 200 个虚拟情境进行模拟(样本量为 50 的虚拟情境不被用于 9 个维度

设计，虚拟情境的数量太少以至于无法达到任何有意义的 D-efficiency）。分辨度Ⅳ总是被采用。在这些设计中，所有双向交互作用都是正交的，排除只有两个维度之间的交互作用（即交互 $X_1 \cdot X_2$）。这种交互作用是通过设计的，与许多其他因素密切相关；因此当它被包含于正交化的因素之中时，通常会阻碍算法的收敛。用灰色强调的栏显示了样本的特性，这是由 SAS 宏为 D-efficient 抽样技术提供的（宏提供的第一个设计被显示；当用不同的随机种子启动宏时，它可能会收敛于一个稍微不同的分辨度，因为算法存在随机性）。

我们可以清楚地观察到，D-efficiency 会随着虚拟情境样本的增加而增加。如果排除案例或虚拟情境全集（维度和水平的数量）增加，它就减少。对不同样本和设计的质量有一个更加直观的认识是通过维度间的平均值和最大相关性，如表 3.3 中间部分所示：为了安全地避免即使是中度相关性（$|r| \geqslant 0.1$），我们必须在 7 个（9 个）维度的设计中使用至少 100（200）个虚拟情境，尤其是当排除不合情理的组合时。[21] 我们也可以清楚地观察到随机样本的代价［请看没有标强调的栏目，它显示了模拟随机样本的平均值，有 10 个（D-efficiency 值）或 1 000 个（相关性和方差）］。在比较两个抽样技术时，D-efficient 的样本更有效。在小部分情况下（例如从一个 $N_u=432$ 个虚拟情境的全集中抽取 50 个虚拟情境或从 $N_u=2\,592$ 个虚拟情境的全集中抽取 100 个虚拟情境），D-efficiency 大概是随机样本的两倍（见第 4 列，显示了 D-efficiency 值的比率）。回想一下，二者的比率表明，在 D-efficient 样本中，我们只需要随机样本中一半的调查对象就能达到相同的参数估计精度。[22] 大概 200 个虚拟情境就能让 D-efficient 样本和随机样本间的差异几乎消失。

表 3.3 **D-efficient 和随机设计中维度的 D-efficiency、相关性及方差**

	D-efficiency			不同维度 X_k 之间的相关性				$var(X_k)$均值	
				$\lvert r \rvert$均值		$\lvert r \rvert$最大值			
	D-efficient 抽样	随机抽样[a] (SD)	D-efficient 抽样与随机抽样的比率[b]	D-efficient 抽样	随机抽样[c]	D-efficient 抽样	随机抽样[c]	D-efficient 抽样	随机抽样[c]
7 个维度：$3^3 2^4$ 设计（$N_U=432$ 个虚拟情境）									
$n_s=50$	79.89	39.13 (4.11)	2.04	0.055	0.108	0.099	0.293	0.443	0.430
$n_s=50$ 排除：$X_1=3$ 且 $X_2\geqslant 2$	74.22	27.87 (7.67)	2.66	0.037	0.106	0.122	0.285	0.427	0.408
$n_s=100$	97.27	78.68 (0.17)	1.24	0.018	0.071	0.050	0.190	0.434	0.430
$n_s=100$ 排除：$X_1=3$ 且 $X_2\geqslant 2$	90.42	72.18 (1.60)	1.25	0.010	0.067	0.048	0.184	0.427	0.408
$n_s=200$	99.38	93.02 (0.56)	1.07	0.010	0.042	0.030	0.113	0.431	0.430
$n_s=200$ 排除：$X_1=3$ 且 $X_2\geqslant 2$	91.46	85.82 (0.55)	1.07	0.010	0.036	0.006	0.098	0.425	0.409

续表

	D-efficiency			不同维度 X_k 之间的相关性				$var(X_k)$均值	
				$\|r\|$均值		$\|r\|$最大值			
	D-efficient 抽样	随机抽样[a]（*SD*）	*D*-efficient 抽样与随机抽样的比率[b]	*D*-efficient 抽样	随机抽样[c]	*D*-efficient 抽样	随机抽样[c]	*D*-efficient 抽样	随机抽样[c]
9 个维度：$3^4 2^5$ 设计（N_U=2 592 个虚拟情境）									
n_s=100	84.31	52.39（1.66）	1.61	0.025	0.079	0.089	0.235	0.439	0.435
n_s=100 排除：X_1=3 且 $X_2 \geq 2$	79.10	47.60（4.42）	1.66	0.023	0.078	0.059	0.232	0.430	0.419
n_s=200	97.13	79.55（0.55）	1.22	0.014	0.054	0.037	0.161	0.435	0.435
n_s=200 排除：X_1=3 且 $X_2 \geq 2$	90.84	72.85（1.48）	1.25	0.012	0.078	0.037	0.234	0.427	0.419

注：a. 10 个随机样本的均值。
b. *D*-efficient 抽样的 *D*-efficiency 除以随机抽样的 *D*-efficiency。
c. 1 000 个随机样本的均值。

有趣的是,随机样本中的维度是高度相关的。最大的相关性(通常表示1 000个随机样本模拟的均值)通常达到大概$|r|=0.2$或0.3的中等效应。因此,许多在虚拟情境全集中的实验因子的独立性,以及因素调查实验内部效度会丢失(关于效度的额外的信息在第6章第2节呈现)。最后,最后两列显示了*D*-efficient样本的维度总是表现出更高的方差[显示了均值方差,在所有维度上的平均(方差的随机抽样平均超过1 000模拟)]。与随机样本相比,更大的方差表明*D*-efficient样本识别虚拟情境维度的真正影响具有更大的统计检验力。此外,我们应该记住,*D*-efficient样本在识别全部有意义变量的影响方面有更大的置信度。

回到确定足够样本量的一般步骤:第一,我们可能要测试大概100至200个虚拟情境的样本来确定*D*-efficiency。当效率的增加超过了调查成本的增加时,才应该使用更大的样本。在我们的经验中,*D*-efficiency的值超过90时,通常只观察维度之间的边际相关性($|r|\leqslant 0.05$),这在社会科学的大部分研究中是可接受的。

第二,要确定有意义的样本量,显示给每一个调查对象的虚拟情境数量n_d是重要的。一方面,大量的虚拟情境可能会使调查对象的负担过重,导致疲劳效应。另一方面,我们必须考虑预算的限制。一般来说,招募更多的调查对象比向单个调查对象呈现若干虚拟情境贵得多。对n_d的一个适当让步取决于虚拟情境的复杂程度。若只有几个可变维度的虚拟情境,调查对象在重复估计相似的虚拟情境之后,很快就会觉得无聊。复杂而冗长的虚拟情境文本,在许多维度上也存在较高的疲劳风险。有关方法研究表明,包含理想数

量7(加减2)个可变维度的虚拟情境,需要在每个研究对象中使用不超过10个虚拟情境。如果使用更多的虚拟情境,研究对象会出现疲劳的迹象(对于调查对象特定数据分析,需要大量的案例,但是我们通常不推荐这个过程;详见第5章)。此外,我们必须考虑到每个区组中虚拟情境的数量也限制了总体的样本量。虚拟情境的总体样本量(n_s)应该是每个区组中虚拟情境数量(n_d)的倍数。例如,每个层块有8个虚拟情境,我们的样本总量就应该是80、88或96个,而不是100个,因为这个样本总量不能够被分成层块预期的样本大小。

第三,我们必须确保样本中每一个虚拟情境能由几个调查对象评估的。若一些虚拟情境没有被评估时,设计中的*D*-efficiency就会被扭曲,某些参数可能无法被识别。如果只有一个调查对象评估单个虚拟情境,那么这个虚拟情境或实验条件就会与调查对象的特征完全混淆。因此,应该有几个调查对象对相应层块的同一个虚拟情境进行评价。一般而言,从全集中抽取的虚拟情境的分部越小(虚拟情境样本的效率取决于每个被评价的虚拟情境),调查对象的差异性越强,那么对每个虚拟情境(层块)进行评价的调查对象就越多。我们建议每个虚拟情境的层块至少分配5个不同的调查对象。但是文献中提供了更加保守的建议。例如在选择实验的相关方法中,建议每个层块至少要有50个不同的调查对象(Bennett & Adamowicz, 2011; Hensher et al., 2005)。但这些数量都是建议,而不是出于统计理论。

选择虚拟情境数量的第四个标准是调查时间。在我们的经验中,对于8个维度的10个虚拟情境(在一般人群中调

查收入公平的例子)反应时间中位数是3.5分钟左右。包括更多维度和/或虚拟情境的更复杂的模块需更多调查时间，如表3.4所示。另外我们的数据表明，调查对象必须在1分钟左右完成对第一个虚拟情境的评估(在收入公平的例子中8个维度的估计量，除了在因素调查模块前的介绍说明可能需要1或2分钟)。在第一个虚拟情境之后，调查对象的回答速度加快。大概在第三个虚拟情境之后，调查对象需要不超过30秒的时间来评估单个虚拟情境。然而调查对象之间也存在很大差异，应答的时间可能很大程度上依赖于因素调查模块的主题和所使用的问卷。因此，如果提前预估应答时间很重要(例如计算调查成本)，那么我们建议对因素调查进行应答时间的预测试。

表3.4 虚拟情境和维度数量(分钟)的应答时间中位数

	维度数		
虚拟情境数(n_d)	*5*	*8*	*12*
10	2.85	3.52	3.85
20	4.97	5.72	6.33
30	6.87	8.37	9.55

注：9个调查分部的每一个的适度应答时间是根据至少$n=104$调查对象和$n=2\,890$虚拟情境评估得出的。
资料来源：德国2009年的普通人口调查(Sauer et al., 2011)。

此外，我们应该意识到，如果希望对调查对象的子样本进行统计测试，就必须确保每个子样本中有充足的调查对象。但总存在一些不回答的情况。因此当计算虚拟情境样本量时要考虑应答率(Champ & Welsh, 2006)。[23]

最后，更精准地确定理想样本量的方法是考虑参数估计精度的统计学理论和计算参数估计的统计检验力。然而这

个过程涉及使用先验的参数估计，以及对参数估计的统计背景和置信水平更深的反思。我们将在后面提供一些信息。对计算统计力的统计细节不感兴趣的读者可以跳过详细的解释部分，在本小节最后的小结部分会给出最重要的建议。

统计检验力计算

统计检验力指当这个假设是错误的时候，拒绝零假设的概率（通常情况下，在自变量和结果变量间没有关系）(Cohen, 1988)。对于给定的显著性水平 α，较高的统计力 b 与以下标准有关：(1)大量的虚拟情境估计；(2)调查对象和虚拟情境特征之间较低的协方差；(3)虚拟情境样本的较高 *D*-efficiency(Hensher et al., 2005; Louviere, Hensher, & Swait, 2000)。统计检验力的文献为简单的随机样本提供了若干个公式，通过计算最小的样本量 n，以达到一个特定的准确水平。这些公式是非常直接的比率，因为可以预先计算出给定比例的标准误（即只要假设所期望的比率值，不需要进一步的信息）。我们可以在一个比例估计的正确度上估算最小的案例数 n（例如，被评为公平的虚拟情境比例），通过使用以下公式(Hensher et al., 2005:185)：

$$n \geqslant \frac{(1-p)}{pt^2}\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]^2 \qquad [3.5]$$

其中，p 是真实比例，t 是介于估计值($\hat{p}$)与真实参数 p 之间的可接受的偏差，$\Phi^{-1}(\cdot)$是标准法线的逆累计，$1-\alpha$ 是估计的置信水平，最大概率$|\hat{p}-p|$分散成多于 t 个百分点，成为了 α。例如，如果我们想要估计虚拟情境的比例，虚拟情境

被评为公平的最大偏差是 4 个百分点，置信水平为 0.95%，并假定真实的比例是 10%，虚拟情境评价的最小数量的为

$$n=\frac{1-0.1}{0.1\cdot 0.04^2}\left[\Phi^{-1}\left(1-\frac{0.05}{2}\right)\right]^2=5\ 625\times 1.96^2=21\ 609。$$

也就是说，如果每个调查对象要评估 10 个虚拟情境，那么至少要访问 2 161 个调查对象。

文献为其他参数提供了相似的统计检验力公式，例如均值。然而，我们必须先知道参数值和标准误。我们可能通过预测试数据来获取这些值。所有这些公式源于简单随机抽样的假设(simple random sampling, SRS)，即只有当每个调查对象评估一个虚拟情境时，它们才是正确的。[24] 一般而言，每个调查对象都会产生更多的判断，导致聚集或分层数据结构。调查对象的特征在第二水平(L2)，同时虚拟情境变量和判断在第一水平(L1，详见第 5 章)。为了获得这种现象如何影响适度样本容量的计算的更多信息，设计效应(*deff*)是有帮助的。这一效应的定义如下：

$$deff=\frac{\text{给定设计下抽样方差}}{\text{SRS 下抽样方差}} \qquad [3.6]$$

如果 *deff* 的值大于(小于)1，设计的有效性比 SRS 更少(多)。为了达到与 SRS 相同的效率，所需的样本量是 SRS 样本容量乘以 *deff*(例如，*deff* 为 1.2，即为了达到与 SRS 相同的效率，我们需要增加 20%的虚拟情境)。第一个经验法则是，当验证 L1 变量的假设时，L1 样本的总量(即虚拟情境的数量)最影响效率。相反，在验证 L2 变量时，L2 样本的总量(即调查对象的数量)更为关键。

更确切地说，对比例的设计效应或对总体均值的估计

(例如虚拟情境判断的均值)是由下面的公式给定的。其中,n_d 表示每个 L2 单元中案例的数量(即每个调查对象中虚拟情境的数量),ρ 是组内相关(这表示调查对象的变化解释了结果变量变化的比例 Y_{ij},不熟悉这些多水平分析术语的读者可先学习第 5 章):[25]

$$deff = 1 + (n_d - 1)\rho \geq 1 \quad [3.7]$$

ρ 值虽然不能提前知道,但可以从其他研究或预测试中使用先验值来获得一般性估计。上面的公式表明,在统计意义上(忽略调查成本),更有效的设计是那些单个调查对象只对若干虚拟情境进行评估(即低 n_d 值),特别是在调查对象在他们的判断原则上有很强的差异(即高 ρ 值)的情况下。

在因素调查研究中,人们通常会对虚拟情境维度的效应和它们在调查对象间的变化的假设感兴趣。因此通常使用回归分析。伴随着这种方法出现的是,在两个水平间(调查对象和虚拟情境变量)、在参数估计间(回归系数和方差组分)、在回归方法间(如,固定的或随机的效应回归)$deff$ 值会不同。相关样本量通常只能通过模拟研究来估计(见 Kreft & De Leeuw, 1998)。那些对统计检验力计算感兴趣的读者可以参考多水平分析的文献(如,Kreft & De Leeuw, 1998; Snijders, 2001, 2005; Snijders & Bosker, 2012)。在本书中,我们只提供一些一般性的建议(这些建议是根据 Snijders, 2001, 2005 改编的)。

为了估计 L2 变量 Z_1 的回归系数(调查对象变量),假设这个变量与其他解释性变量 Z_m 无关,同时这一线性添加模型是正确的(例如方程 3.1 所示的模型),可以用设计效应

等同的公式作为报告比例和总体样本均值的方法(见方程3.7)。[26]也就是说,与SRSs相比,这些回归系数总是带有较低的检验力。[27]

反之,相较于SRS, L1变量 X_1(虚拟情境维度)的回归系数估计的检验力更多或更少(Snijder, 2005)。如果以下三个假设都成立,那么设计效应会更低,检验力会更高:(1)L1变量与任何其他解释性变量 X_k 都不相关(在不排除不合逻辑或不合情理的案例的情况下,这对于所有好的实验设计来说都是为真的);(2)对于所有L2单元(调查对象;将虚拟情境分成块且采用了 D-efficientt 抽样技术),变量 X_1 都有一个恒定值;(3)在回归系数中不存在调查对象间的差异(即没有随机斜率;同样,不熟悉多水平分析术语的读者可以先学习第5章)。如果这三个假设都正确,那么设计效应的计算如下(参阅Snijder, 2005:4):

$$deff = 1 - \rho \leqslant 1 \qquad [3.8]$$

然而,这个公式只有当回归系数中不存在调查对象间差异时才适用。用不太技术化的方式来说,只有当调查对象在他们的判断原则中显示了实质性社会共识时才是正确的,这种情况更可能出现在同质性的调查对象样本中(例如来自同一年龄层、有着相同教育水平的调查对象)。在这些情况下,每一个调查对象若干个虚拟情境的设置就提高了统计检验力,这通常涉及调查个体内估计的更高的检验力。但是在大多数应用情况中,更现实的场景涉及调查对象间的变异(即模型有随机斜率)。在这样的情况下,对于 $deff$ 公式,相应的检验力的计算会更加复杂。X_1 的设计效应也取决于L1

的组内方差(表示 s_{x1}^2,并假设在所有组/块上都是常数),L1 的残差 σ^2, L2 的残差 τ^2,随机斜率的方差为 τ_1^2。设计效应可以大于或小于1。换句话说,统计检验力可以比 SRSs 的更高或更低(参阅 Snijder, 2005:4):

$$deff = \frac{n_d \tau_1^2 s_{x1}^2 + \sigma^2}{\tau_1^2 s_{x1}^2 + \tau^2 + \sigma^2} \overset{?}{\gtrless} 1 \quad [3.9]$$

上面的公式表明,尤其是对每个调查对象 n_d 面对大量虚拟情境的情况,设计效应会随着随机斜率 τ_1^2 的方差增加而增加(因此统计检验力会减少)。如果调查对象之间存在一个较低的社会共识(相应地随机斜率的方差较高),调查对象数量的增加比单个调查对象评估数量的增加可能更有效地提高 L2 和 L1 的变量的效力。我们应该衡量效率的提高,而不是增加调查对象和调查费用。

我们通常会对跨水平的交互作用感兴趣(即在调查对象和虚拟情境变量之间的交互作用)。由于所涉及的因子数量,很难精确地确定这些交互作用的统计效力。这对于因素调查而言尤其正确。但是在我们的经验中,通过使用方程 3.7 计算出的跨水平交互作用损失的检验力通常可以很成功地接近 L2 变量的总体均值和主效应。可以通过软件包来计算更为精确的跨水平交互作用的统计效力和其他分层模版中的参数值。[28]有意义的跨水平交互作用应该减少在回归系数中调查对象间的变异,从而增加 L1 变量估计的统计效力。然而,不相关的交互作用减少了自由度的数量,因此减少了统计力。

本节中提供的所有公式都是基于这样的假设,即解释性

变量之间互不相关。如果这些变量是相关的(至少对于调查对象的变量是现实的),那么就不存在公式;但你可以使用斯尼德斯和博斯克(Snijders & Bosker, 1993)提供的类似产品。[29]

在一般意义上,这种讨论表明所有忽视分层数据结构的数据分析对于显著性水平和统计力的计算都是具有误导性的。因此,为了正确分析每个调查对象要进行的多种虚拟情境评估,我们必须使用一些针对分层数据结构的分析技术,例如多层回归或带有稳健标准误的回归(详见第5章)。

总而言之,除非有模拟研究支持因素调查的样本量,否则最好采取保守措施。为了估计调查对象特征和跨水平交互作用的影响,我们需要大量调查对象来提高统计检验力。特别是在希望获得回答者之间差异高的应用中,回答者的数量(不仅是虚拟情境评估数量)对于提高统计结果的效力很重要。如果每个调查对象都只评估一个虚拟情境,那么就可以计算出相应调查对象数量的上限。因此,案例的数量减少到调查对象的数量,并且适用于SRS的标准公式。所期望的斜率方差越小,从调查对象个体中收集到的一些观察结果的效率损失就越低。换言之,通过更少的调查对象收集更多的虚拟情境评估所减少的调查成本越来越无害(就统计检验力而言)。考虑到认知的限制,例如疲劳影响,上限大概是每个调查对象分配10个虚拟情境。请注意,如前所述,特别是在具有异质性的调查对象样本的应用中,每个层块大量的调查对象(扩大调查对象的数量)将减少混淆调查对象变量中虚拟情境变量和层块影响的风险。

小结

为了确定合理的虚拟情境数量，我们应该考虑 *D*-efficiency。它可以通过不同的大分部实现(一开始大概 100—200 个虚拟情境虚分部来确定效率)；为了避免疲劳和学习的影响，每个调查对象应分配不超过 10 个虚拟情境；虚拟情境分部应该是每个层块虚拟情境的倍数；每个虚拟情境(层块)应该被若干个调查对象估计(至少 5 个调查对象；当调查对象样本异质性的增加时，每个虚拟情境需要的调查对象数量也在增加)。最后的标准是回答虚拟情境模块需要的调查时间(经验表明，调查对象需要大概 30—60 秒估计一个虚拟情境。我们应该对手头的因素调查的调查时间进行预测试)。

足够的调查对象样本量取决于调查对象的差异性(差异性越大，就需要越多的调查对象)、增益影响(对于跨水平交互作用或子集分析，我们需要更多的调查对象)，以及我们期望达到的统计检验力。对于 L2 和 L1 的估计，随着不同调查对象收集到的判断的数量的增加，获得统计检验力也会增加。特别是在异质性的调查对象样本中，建议增加调查对象数量。因为在调查对象判断原则中有很大的差异，降低了探查 L1(即虚拟情境)变量效应的能力。尤其对于那些对评估调查对象变量影响和跨水平交互作用感兴趣的人来说，最好测试更多的调查对象，同时对每个调查对象使用较少的虚拟情境。

可以通过使用文献中提供的公式或软件程序进行多水平分析来估计统计检验力。所有样本量的计算应该考虑可能的无应答和任何有意的子集分析，这需要抽取大量样本。

第 6 节 | 总结:实验设计的清单和工作流程

下面的清单包括建构实验设计的最重要问题。这五个实践步骤的顺序遵循了便捷工作流程的逻辑。

步骤一:维度和水平的说明

- 是否有足够数量的维度来保证调查对象不会感到无聊或遗漏信息?
- 维度的数量是否足够小,以避免调查对象的负担过重(至少对那些年龄过大或认知能力较低的人来说)?建议使用不超过大概 7(加减 2)个维度,尤其是当每个调查对象有若干个虚拟情境时。
- 在维度上是否有类似数量的水平以避免水平数量的影响?
- 水平数量是否允许对非线性影响进行识别?为了确定非线性影响,至少需要三个不同水平。
- 所有水平都是必要的吗?是否有更合理的设计?只有当研究目的明确需要更多数量时,才建议使用三个以上的水平。

- 当使用文本性虚拟情境时，维度和水平结合的文本是否流畅易懂？
- 我们可以避免不合逻辑或不合情理的组合吗？最好的方法是在确定维度和水平时就避免这种情况。

步骤二：排除不合逻辑的案例

- 在实验设计的候选集中是否存在任何需要删除的不合逻辑或不合情理的组合？方法论研究建议至少要删除那些非常不合情理的组合，因为它们很少会出现在真实生活中，而且调查对象可能不会认真对待。

步骤三：案例数量

- 为避免疲劳效应，每个调查对象所面对的虚拟情境是否足够少？建议给每个调查对象分配不超过 10 个虚拟情境。
- 虚拟情境的总量可以被分为每个调查对象或层块中的虚拟情境吗？
- 虚拟情境和调查对象的数量是否能够确保足够的统计检验力，来检测所有效应的影响，包括虚拟情境维度和调查对象特征之间的跨水平交互作用，并具有足够的精度？
- 我们可否预期得到来自不同调查对象对虚拟情境和层块的充分评价，以避免混淆虚拟情境维度和层块与调查对象的变量吗？第一条经验法则是确保每个虚拟情

境被至少 5 个不同的调查对象评价。随着调查对象异质性的增加，每个虚拟情境也需要增加调查对象。

步骤四：设定实验设计

- 设计是否允许识别所有虚拟情境维度和交互作用的影响，而这些影响对虚拟情境评价是不可忽略的？在进行实验设计时，最重要的问题是识别参数。
- 当排除某些组合（因为它们不合逻辑或不合情理）的时候：在进行实验设计时是否考虑到了这个限制？如果在此之后删除了这些组合，一些设计的特征可能会丢失。
- 设计是否允许大量关于虚拟情境维度影响（以及它们的交互作用）的信息，即它是否达到了一个令人满意的 *D*-efficiency 的水平？如果没有，我们应该回到步骤一至三，测试维度和水平数量更少的设计，抽取更多虚拟情境数量，或者更少的限制，例如必须从实验设计的候选对象中排除维度组合。
- 是否打算用非线性模型估计虚拟情境结果，例如逻辑回归或概率模型？是否对更高的参数估计精度（统计效力）有兴趣？如果有，你可以考虑在第 3 章第 2 节第三部分中对类别结果的特别建议。

步骤五：划分层块

- 单个层块允许识别最重要的参数吗（因为它们不会与

层块混淆)?

- 切分成层块为单个层块保留了一个较高的 D-efficiency 值吗?如果想要估计调查对象的特定参数,例如随机斜率,这些方面尤为重要。

步骤六:总论

因为在目标标准间有许多权衡取舍,所以我们必须经常重复之前的步骤。例如,就统计效率的增益(或损失)而言,建议对不同的设计规格(不同数量的维度、水平和虚拟情境,不同设计方案)估算最大的 D-efficiency 值,以观察设计中微小调整的效果。然而,所有的决策都应该以有意义的参数的可识别性这个核心标准(包括在虚拟情境维度间的交互作用以及可能的非线性影响)为导向。因此,这个标准应该是最后的检查:设计是否允许所有具备令人满意水平的精度(统计效率)的重要参数被识别?我们可以通过在虚拟情境维度(和层块)之间使用相关矩阵来评估这一问题,包括评估多重共线性模式,或模拟一些回应的数据来决定是否能估计所有效应。

第4章

调查设计

本章描述了在实验设计完成后如何实施因素调查的调查模块。大部分步骤和其他调查研究一样,但也有一些针对因素调查的不同特点。例如,在因素调查中,一个调查对象要面对几个虚拟情境。因此调查对象必须反复评估非常相似的对象或情况。当评估多种虚拟情境时,若呈现的反应量表不足以测量调查对象的反应,这会导致应答删失(censored response)的风险,因此需要更多的反应量表或模版来解决。此外,对于因素调查,我们必须准备和分配比通常情况下更多的不同问卷版本。接下来,我们会提供实用建议,先从足够的调查对象样本(第 4 章第 1 节)和回答量表(第 4 章第 2 节)开始,然后针对文本、表格、图片或录像的虚拟情境的选择和呈现的模式,以及虚拟情境和维度排序的规范(第 4 章第 3 节),接下来讨论了调查模式(第 4 章第 4 节)和实施大量不同虚拟情境和问卷版本技术的实现方法(第 4 章第 5 节)。最后,提供了给调查对象的充分指导的一些实用建议(第 4 章第 6 节),并预测试了因素调查模块(第 4 章第 7 节)。

第 1 节 | 调查对象样本

是否需要对目标群体的代表性样本进行调查，或从便利抽样中得出结论？解决这些问题的标准是内部和外部效度，以及获取调查对象所需的资源。

内部效度是指结果变量的变异是否真的由实验干预引起的(Aronson, Wilson, & Brewer, 1998; McDermott, 2002)：对实验刺激(虚拟情境)的操纵是否导致了观察反应，或是还有其他任何没能控制的虚假因子？为了获得因果结论，就需要较高的内部效度。在实验中安排参与者随机接受实验刺激(虚拟情境)以确保内在效度，这是所有实验的关键环节。因此，最重要的是确保这些虚拟情境刺激(虚拟情境层块)完全随机地分配到调查对象，并有足够数量的调查对象对每个单独的虚拟情境进行评价，以避免调查对象特征与虚拟情境特征的混淆(其他对内部效度的威胁来自认知负荷和其他方法论的问题，例如调查对象误解任务或疲劳；我们将在第 6 章第 2 节回到这个主题)。

方法论研究表明，当调查研究的数量被限制在一个适度的 7(加减 2)个变量维度和 10 个虚拟情境内，即使是年纪较大或受教育程度较低的调查对象，也能提供高水平的应答一致性，使得内部效度没有超负荷认知或疲劳效应(Auspurg et

al., 2014; Sauer et al., 2011)。然而随着复杂水平的提高，年纪较大、受教育程度较低或不太熟悉主题的调查对象更可能产生不一致的回应。此外，面对大量的虚拟情境和维度，调查对象倾向于忽略一些维度的简化启发式（simplifying heuristics）方式（Auspurg & Jäckle, 2012; Sauer et al., 2011）。这两种反应模式都可能会降低内部效度，因为调查对象可能没有意识到一些实验的刺激。

出于这些原因，当处理调查对象的异质性样本时，遵循有关虚拟情境模块复杂性的方法论建议尤为重要。否则很难觉察出判断或决策规则的不同，在子群体差异方面得出错误结论的风险也很高。一个可能错误的结论是，一些调查对象群体对一些维度的重要性评估低。而事实上，不是他们不重视这些维度，而只是没有注意到。调查对象的任何认知负担都可能降低内部效度和数据的统计效力（因为判断错误增加了错误方差，导致了统计检验力的降低）。因此，正如一些研究者所建议的，当设计调查实验时，我们不仅要考虑统计效率，还要考虑“调查对象的效率”（Ferrini & Scarpa, 2007; Louviere, Islam, Wasi, Street, & Burgess, 2008）。当调查对象是专业人士或学生时，可能需要采用更复杂的虚拟情境模块，以避免产生“调查对象的效率”的问题。在某些情况下，基于研究目而使用了多于7（加减2）个维度，与理论相关的重要权衡和维度间的交互作用可能无法得到解决。在这种情况下，在对因素调查模块进行预测试时，我们应该特别注意认知负荷的迹象。

外部效度是指能将结果推广到其他不同样本、环境、测量或实验干预（Mutz, 2011）。调查对象的异质性样本能确

保这种推广性吗？因此因素调查的调查对象样本群体很不同。许多调查依赖于大规模的人口样本，而许多其他研究使用学生样本或特殊群体，例如专业人士(Wallander, 2009)。

如前所述，所有实验的主要优势是验证因果理论。在验证因果关系时，一个重要的起点是剔除真实世界中的复杂性而关注单一的因果机制。所有的实验室实验都抹掉了这种真实的复杂性。它们的目的不是对行为进行普遍化，而是验证行为背后的机制(Aronson et al., 1998; McDermott, 2002)。总的来说，因素调查的实验效度并不依赖于使用人口抽样。在那些希望验证出因果机制的案例中，这个目标可以通过实验室的便利抽样实现。[30]然而，如果这些机制被调查对象特征进行调节(即如果在调查对象变量和虚拟情境变量之间有交互作用)，那么从调查对象样本推广到其他群体就会出现问题。在同质性样本中，很多变量是常量(通常是有意设置)，这样才能排除这些调节性变量的机制。

验证理论的下一个步骤是增加复杂性和交互作用，来达到更高水平的理论理解和更高的外部效度(McDermott, 2002; Mutz, 2011)。调查实验相对实验室实验有一个优势，即能够以相对较低的投入研究异质性较强的调查对象群体。调查研究很容易地超越地理位置限制，抵达许多不同的调查对象群体，或精确达到研究者希望推广的调查对象群体(Mutz, 2011)。因此，因素调查研究对于了解特定职业群体的专业判断和信念特别有帮助，例如作为管理者、青少年法庭的法官，或实验室实验很难招募到的雇主(述评见 Wallander, 2009)。此外，它们还有助于研究差异大的社会规范和态度，揭示这种子群体差异是因素调查研究的主要优势之

一(Byers & Zeller, 1998; Rossi & Anderson, 1982)。

因此,理想的调查对象样本总是取决于具体的研究目的,为了验证被认为是普遍性的因果机制(例如利他主义或理性主义),我们最初可能会选择同质性的(学生)样本。但如果想了解对所有人的影响,那么对于描述性统计或估计所有可能接受因素调查刺激实施中的人的平均实验效果,一个全国性的随机样本会更合适(Mutz, 2011)。尽管因素调查研究不要求必须使用异质性样本,但采用这种样本是因素调查方法的另一个主要优点(详情见第 6 章第 2 节)。

我们在思考调查对象样本适切性的时候,必须考虑研究资源,大学生通常是最容易获得的便利样本。更多的异质性调查对象样本会引发虚拟情境估计中调查对象内部差异的变化。单一虚拟情境维度的纯影响的统计效力更小(见第 3 章第 5 节第二部分)。研究者需要决定是选择对虚拟情境维度影响了解更多[在这种情况下,尽可能控制许多干扰(调查对象)变量是令人满意的],还是更深入理解社会群体在判断中的差异。后者允许更广泛的推广度,但是用于检测单一虚拟情境维度影响的统计效力减弱了。

关于抽样方法(例如随机或定额样本),我们可以参考一般的调查文献(例如 Groves et al., 2009)。因素调查在抽样方法上没有什么特异之处,但是当实施调查时,我们应该记住,虚拟情境必须完全随机地分配给调查对象,否则会损害实验的内在效度。

第 3 章第 5 节提供了计算调查对象的样本量的方法。外在效度不仅依赖于调查对象样本,也关乎虚拟情境的现实程度以及推广度。在第 6 章第 2 节中会更详细地讨论这些问题。

小结

只要遵循关于复杂性的方法论建议，就没有认知负荷的严重风险[即不超过大概 7(加减 2)个维度和 10 个虚拟情境]，因素调查可以适用于一般人群样本和年龄较大、受教育程度较低的调查对象。一般来说，特别是当研究目的是验证理论和因果机制(对于因素调查尤为适合)时，对内部效度的关注要优先于外在效度，因此向调查对象随机分配虚拟情境就非常关键。因果结论不一定依靠从普通人群中抽取随机代表性样本。我们也可以使用方便抽样，因为这在几乎所有的实验室实验中很常见。

然而，更广泛的推广度需要更广泛的调查对象，只有调查对象变量中的变化能确定这些变量是否影响了因果机制。因素调查的主要优势之一是可以很容易地推送到研究者希望推广到的异质性群体。因此，在研究子群体的差异或是特殊(专业)群体的判断原则时，因素调查优于实验室实验。研究者在呈现他们的因素调查结果时可以利用这一特点。然而我们也注意到，更多异质性的调查对象样本需要大量的调查对象来达到相同的估计维度影响的统计效率。增加对子群体间差异的认识也牺牲了实验刺激(虚拟情境维度)单一影响的部分效力。

第2节 | 回答量表

对虚拟情境的回应构成了研究的主要结果变量，进而也影响了因变量的测量属性和数据分析技术（见第5章）。根据杰索的研究（Jasso, 2006），回答量表可以被定义为一系列有序类别（例如，收入公平的例子）、一系列无序类别（例如，对犯罪行为的不同判决）、金额数量（例如，收入公平），或概率（例如，以特定方式行动的概率）。

迄今为止，有序等级量表是最常用的回答量表的类型（Wallander, 2009）。这种量表最大的缺陷是应答删失的风险。一旦调查对象选择量表的末端分数，他们就不能再为剩下的虚拟情境选择更极端的类型或者分数。此外，他们必须完善自己的判断（Jasso & Rossi, 1977:643），同时他们不太可能以同等间隔的方式使用量表（Louviere, 1988）。

为了解决这些问题，一些研究者建议使用量级回应量表，这是由史蒂文斯（Stevens, 1957）首创的，特别是数字匹配技术（例如，见 Jasso, 2006）。调查对象被要求匹配一个数字（或标记一条线的长度），每个刺激（虚拟情境）的感知量或强度是成比例的。图4.1提供了这种技术的一个示例。调查对象需要通过选择任一数字来表达他们所感知到的不公平程度，用负数（正数）标识他们认为的不公平低（高）收入，而

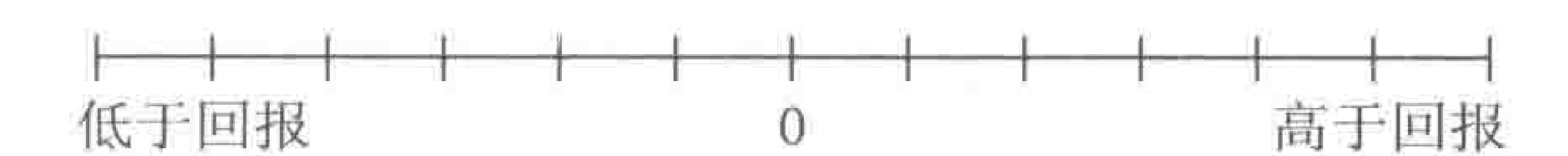
零则代表他们认为收入是公平的。调查对象可以使用他们不受限制的连续数字，这样避免了受限的问题。因此，我们期望在比率尺度上达到一个衡量标准（Lodge & Tursky, 1981）。为了激活这个连续体，指示中应该明确地提到使用分数和小数的可能性（Jasso, 2006）。所有的虚拟情境评估都被计算为相对评估，第一个刺激（虚拟情境）必须标准化，成为所有调查对象的一个参考点。[31]

我们希望你用数字呈现你的判断。

零代表完全公平，负数代表低于应有回报程度，正数代表高于应有回报的程度。越高或越低于应有回报的程度，你选择的数字的绝对值就越大。高于或低于应有回报的程度越轻微，呈现的数值就越接近零；反之，高于或低于应有回报的程度越高，呈现的数值就越远离零。

例如，－82 表示这个人低于应有回报的程度是得分为－41 的人的两倍。

公平的评价量表可能如下所示：



对于每个描述，请写下最符合你对员工每月总收入公不公平判断的数字。你可以使用的数字范围是没有限制的。一些调查对象喜欢从－100 到＋100 绘制个人量表，而其他人则喜欢更大或更小的范围。你可以选择任何实数，包括小数和分数来呈现你的判断。

1 号雇员：

一个 45 岁的人……（进一步的虚拟情境文本）

你的判断：________

注：这个例子改编自 Jasso(2006：415)。

图 4.1 量表的案例

然而，使用量级量表有很多缺陷。如果调查对象没有对相关的虚拟情境进行评分或者评分为零，那么这个调查对象的所有虚拟情境评估都是无效的（因为零的划分是没有定义的）。因此人们必须识别出一种可能所有对象都认为是不公

平的相应的虚拟情境。我们还应考虑对相关虚拟情境的强制性反应(但是这有触发不愿意完成这项任务的调查对象完全退出的风险)。另一种选择是让研究者提供一个参照评分,但这种选择又限制了调查对象使用自己的连续数字。方法研究已经表明量级量表(一般情况和尤其针对因素调查的情况)比起另一种评分量表,能产生更高比例的无反应项目,至少在它们被用于自我管理(self-administered)的调查时。许多调查对象倾向于通过评价虚拟情境是公平的来跳过量级评估,或者产生非常极端的答案。这些答案很可能代表他们的抗议。如果把这种回答囊括在内,可能导致对收入公平的评估超出合理结果区间(Auspurg et al., 2014; Schaeffer & Bradburn, 1989)。[32]

出于这些原因,我们建议使用更方便的反应量表(例如评级量表)。为了避免应答删失的情况,以及为调查对象提供足够的自由来区分虚拟情境,我们应该通过比普通项目量表更多的应答类别。我们的经验表明,调查对象在包含11个类别的量表中表现出色,反之,当提供更多的类别(例如范围从0—100)时,调查对象使用较少应答类别(例如10、25或50个类别)。证据表明,年纪较大且受教育程度较低的调查对象在使用量表的方式上尤其会遇到问题(Sauer, Auspurg, Hinz, Liebig, & Schupp, 2014)。告知调查对象虚拟情境可以按照任何顺序进行评分且可以改变等级,可进一步避免应答删失。应该允许调查对象反复斟酌并纠正他们的判断。另外,通过随机分配给每个调查对象虚拟情境,我们可以避免应答删失的风险(在评估进程中增加)与因素调查模块最后呈现的虚拟情境之间的相关性(关于虚拟情境排序

的更多细节，见第 4 章第 3 节）。

等级的方法虽然很少在因素调查研究中使用，但我们也可以采用。这些方法要求调查对象在特定结果中对所有虚拟情境进行排序。这种方法还包括配对比较法，要求调查对象在同时呈现的两三个情境中做出选择（例如，见 Buskens & Weesie, 2000）。等级的方法还可以避免调查对象的应答删失。而且，调查对象还被迫区分虚拟情境案例，这可能会防止反应定势，从而提供更多关于判断原则的信息。然而当调查对象必须对他们认为无关紧要的虚拟情境进行排序时，这种策略也会带来随意的回答（例如，他们被强迫对他们认为同样水平的不公平虚拟情境进行排序）。关于调查的复杂性，研究者还未有一个令人信服的结论，即调查对象更清楚哪种类型的任务，对虚拟情境是评分还是排序（Alriksson & Öberg, 2008: 249）。但是，要注意到排序任务常常在数据分析中受到更多限制（例如，我们需要一个更大的观测最小值来估计逻辑或概率的回归；我们通常只通过跨水平交互作用来评价调查对象样本的影响，即使我们只对他们的主效应感兴趣）。对这些任务感兴趣的读者可以查阅第 6 章第 1 节，那里讨论了通常采用了这些回应的任务的相关方法（即联合分析和选择性实验）。

通过直接提问也可以避免应答删失（例如，员工公平收入的金额在虚拟情境中的呈现）。类似地，调查对象会被问及关于金钱数量的问题，即他们为虚拟情境中提供的产品或选项的支付意愿，或是会被问及当面对虚拟情境中描述的场景时他们做出特定行为的可能性。这种直接的提问方式可以避免锚定效应。锚定效应存在于虚拟情境中应答的类

别或(收入的)数量中,进而导致调查对象评估的偏差(Markovsky & Eriksson, 2012)。[33]尽管如此,对于大对数应用,我们并不建议执行这种应答任务。对调查对象而言,直接指明数量可能比在不同应答类别中进行选择要求更高。在我们的经验中,对于这些任务,我们要预计到丢失值和反应定势会更高。此外,对因素调查进行的部分间接评估工作也会因为这种评估而丧失。因此,社会期望的偏差可能会增加。在目前的研究现状下,合理使用评价量表在调查研究和因素调查研究中具有较高的实用价值。一种预防锚定效应的方法(在一些研究中有指出,例如 Schrenker, 2009)可能会指示调查对象,虚拟情境中提供的值是随机分配的,不一定会反映"真实世界"(Jasso, 2012)。

研究者已经有了将多重应答量表加入单一虚拟情境的积极经验(例如见 Abraham et al., 2010)。与其他典型的调查一样,我们必须保证多重应答任务不会过于复杂(进而导致调查对象负担过重),从一个量表到另一个量表中没有很强的延滞效应,并且应答时间是在因素调查模块的限制范围内(关于应答时间,详见第 3 章第 5 节第一部分)。预测试中可以探索这些问题,也可能使用虚拟情境来引导质性访谈。在访谈时,调查对象会被要求解释他们的判断。研究者可能会用这个故事接下去发生的虚拟情境来激发进一步的讨论[Ganong & Coleman, 2006;为因素调查中这个特定的变量杜撰新术语——多部分因子虚拟情境设计(multiple segment factorial design),见第 7 章第 1 节]。

另外关于应答量表的建议是使用对研究目的而言有意义的应答维度,且提供如何使用的具体说明。当使用评分量

表时，最好在端点和中间进行标记以保证所有的调查对象对应答任务有相同的理解(Louviere, 1988)。还可以考虑在因素调查模块的一开始使用一些培训性质的虚拟情境，在我们的经验中，即使是第一个虚拟情境(training vignettes)，也产生了有效的判断(只针对较长的反应时间，这些判断显著区别于更高级的判断，因为学习任务需要更长时间)。但是为因素调查任务提供培训就依赖于调查对象的因素了，例如与话题和调查相关的经验，以及模块的复杂性。

总之，在选择应答任务类型时，研究者可能会使用适合他们研究目的的任何量表。但是基于因素调查研究的悠久传统以及我们对其利弊的讨论，使用等级量表是一个合理的建议。排序的方法或匹配比较为等级量表提供了另类最佳选择，详情见第6章第1节。

小结

我们应该仔细考虑应答量表的选择。它的值代表了主要的研究结果，定义了数据分析的统计模型。定序类别是等级量表中被最广泛使用的，但是我们也采用名义类别或是对数量和概率的直接提问。当调查对象必须评价若干虚拟情境时，所有封闭式的等级量表具有引发应答删失的风险。出于这样或那样的原因，一些研究者推荐量级应答量表，尤其是数字匹配技术。然而这些量表存在许多严重缺陷(例如，缺失值或异常值的比例很高，这会导致不合常理的结果)。因此我们建议不要使用它们。预防应答删失风险和结果的最好方式是提供足够数量的应答类别来区分虚拟情境，以及

允许调查对象改正他们之前的判断。此外，我们应该对虚拟情境随机排序，在所有虚拟情境中平均分配应答删失的风险。直接提问数量或可能性可以避免应答删答和锚定效应，但是这种方法与高比例的缺失值相关联，容易产生社会期望偏差。

因此，对于大多数应用而言，我们建议使用大约有 11 个应答类别的标准等级量表。

第 3 节 | 呈现模式

文字或表格

迄今为止，几乎所有因素调查的应用都是以文字方式来呈现的虚拟情境。另一种选择是在表格中呈现虚拟情境（见图 4.2），通常用于相关多因素调查研究（例如，联合分析和选择性实验，见第 6 章）。使用文字虚拟情境的一个理由是，虚

1 号雇员：

教育水平：	职业训练	年龄：	50 岁
工作：	社工	性别：	女
工作经验：	很多	子女：	3 个
工作任期：	很短	净收入：	每月 2 500 欧元

这个人的月收入公平吗？是太高还是太低了？

不公平 太低了					公平					不公平 太高了
−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5
■	■	■	■	■	■	■	■	■	■	■

资料来源：我们的例子。

图 4.2 虚拟情境的图表呈现案例

拟情境中的短篇故事可能增强调查对象对情境的理解，并有助于想象自己身处这些情境中，或与被描述的人物产生共情。文字形式可能会减少社会期望偏差，因为比起在不同表格单元中的突出显示形式，调查对象不太可能被嵌入运行文本的单一维度所吸引。

表格式的虚拟情境减轻了维度的顺序效应。如果评估被呈现的维度顺序所影响，这种效应就会出现，例如在虚拟情境文本开始和结束时维度的位置会吸引更多的注意力（参见文献中的首因效应和近因效应；见 Krosnick & Alwin，1987；Schuman & Presser，1981）。对大学生进行实验分析表明，这样的顺序效应只存在于非常复杂的任务中（12 个维度和复杂的答复任务；见 Auspurg & Jäckle，2012）。这个结果支持了只使用大约 7（加减 2）个维度。但是在涉及年纪较大且受教育程度较低的人群时是不充分的，他们可能会受到顺序效应的影响（Schwarz，2007；Schwarz & Knäuper，2000）。因此，在考虑当前的研究状况时，我们可以通过随机调整维度顺序来中和所有可能的顺序效应。相比文字格式，这种方法在表格格式中更容易实现。因为在设计文字虚拟情境时，必须与维度结合起来，使文本读起来尽可能自然和流畅。一些维度（例如虚拟情境中的人物性别和收入）比起其他维度，可能在一开始或结束的文本中更具有逻辑性。这就意味着轮换虚拟情境维度的顺序与设计流畅的虚拟情境文本的目的相冲突。相反，在表格格式中，顺序并不是按照语法方式有序排列的。维度顺序的轮换应该在调查对象间的设计时就实施（即对单个调查对象而言，维度顺序应该保持不变）。否则，调查对象很可能会感到困惑。

因素调查对于国际比较而言是一种很有吸引力的方法，因为与普通项目化问题相比，在虚拟情境中，案例的细节描述可对刺激因素进行更高的标准化。例如，在虚拟情境中包括了工作经验和工作时长的信息，我们可以控制国际女性劳动力供给的差异，导致更直接地只对性别收入差异的公平性比较(JGPGs)。但是我们应该确保在所有语言中都采用相同顺序的虚拟情境维度，否则，国家之间的差异会表现为顺序效应。用表格形式呈现的虚拟情境很容易到达这个目标，因为在流畅文本中，优先的维度顺序可能会因为不同语言而变化。这些语言在句子和段落中运用了不同的逻辑来排列单词和短语。

目前还没有对文字和表格式的虚拟情境进行实证比较的研究。我们对大学生进行的收入公平的因素调查实验的初步分析表明，如果表格式的虚拟情境不是过于复杂的话，会与文字式的虚拟情境产生相似的评价(在实验中有 8 个维度)。因此，我们可能主要基于因素调查的主观应用来增强与“现实生活”评价的对应关系，从而产生心理上的真实性(详见第 6 章第 2 节)。表格式的虚拟情境可能更适合包含决策标准列表的决策性任务(即简历或许多消费产品的描述)。然而与文字格式相比，表格式的报告更容易出现社会期望偏差和实验者需求效应，因为调查对象的注意力明显受限于研究者的操纵。因此，表格式的虚拟情境更直观地呈现维度，有助于调查对象形成更一致的判断(在一个简短的应答时间内)。类似的论证涉及虚拟情境中的不同维度是否应该被标注(例如字体加粗呈现)。这种形式的呈现也可能触发社会期望偏差，因为调查对象会更容易识别出被标注的维度。

虚拟情境的排序

另一种顺序效应可能由于虚拟情境的连续位置引起。前续的虚拟情境和对它们的评估可能会产生一种延滞效应，影响对后续虚拟情境的判断，这在调查研究中被称为“晕轮效应”（halo effect，Tourangeau，Rasiniski，Bradburn，& D'Andrade，1989；Wirtz，1966）。此外，例如学习效应、疲劳效应以及应答删失等方法效应，会在一些虚拟情境的评价后出现（即它们本来就与虚拟情境的排序是混淆的）。调查研究显示，这些效应在认知要求高的任务上表现得更明显。这更突出了以下指导原则对应对不同复杂性问题的必要性。一个可能防止虚拟情境顺序效应的方法是，每个调查对象只评估一个虚拟情境（这可能也减少了社会期望偏差，因为只有在调查对象间设计才会有虚拟情境维度的变化，见第 6 章第 2 节的效度内容）。但是，这种策略也意味着数据收集成本会更高（由于需要更多数量的调查对象来实现特定数量的虚拟情境评价）。另外，要排除所有类型的调查对象内分析（within-subject analysis）。因此，大多数研究者倾向于对每个调查对象使用几个虚拟情境。这种情况下，我们应该确保顺序效应的出现不会受限于特定的虚拟情境。通过轮换虚拟情境就能实现这个目的，也可以中和顺序效应对参数估计的影响（尽管它可能会增加误差方差）。最简单的方法是随机安排调查对象。

轮换和随机顺序比一开始就使用极端虚拟情境案例更可取，这是一种避免应答删失的技术。之所以这样建议的原

因是，从可能引发最极端的反应（例如，因为它们显示最不公平的过高或过低的收入）的虚拟情境开始，有利于校准调查对象的评分端点（例如，见 Garret, 1982; O'Toole et al., 1999）。然而调查对象对极端情况的观点也可能不同；如果是这样，系统的排序将避免一些但不是所有的应答删失。在这种情况下，跨水平交互作用中的参数估计会存在偏差（因为参数估计的效应量是由于调查对象应答删失而出现向下的偏差；详见第5章）。因此，尤其对于子集分析，在问卷开始时就固定放置极端虚拟情境可能是弊大于利。基于上述理由，我们建议对每个调查对象使用随机的虚拟情境排序。

当应答任务是对虚拟情境进行单一排序时，至少在纸质问卷上，建议把虚拟情境分别印在单独的卡片上，然后交给调查对象。这样，调查对象可以给卡片排序，然后由他们自己或者访问员记录这些排序。

图片或视频的呈现

虚拟情境除了以文字形式呈现以外，还能用图片来呈现。计算机辅助访问拓展了视听方式呈现的可能。在穆茨（Mutz, 2010）的研究中可以找到若干个这样的应用实例。一个突出的例子是使用视频虚拟情境操控调查对象关于邻里质量的看法。克里森、库珀、法利和福曼（Krysan, Couper, Farley, & Forman, 2009）进行了一项这样的研究，探究种族隔离的原因。迄今为止，我们还不太清楚白种人和其他主要群体与少数民族分开居住的偏好，多大程度上是由于种族偏见还是其他邻里偏好所造成的。许多非裔美国人和少数民

族居住聚集的地区，通常存在以下特征：居民的社会地位低于平均水平、犯罪率高、公共设施相对较少，以及学校质量低。在现有的邻里关系中，这些特征与居民的种族背景的相关性太强，很难分离出来。我们可以用因素调查研究方法虚拟一些邻里关系，析出某一个影响因素。与文本虚拟情境相比（其应用见 Emerson，Yancey，& Chai，2001；John & Bates，1990），视频虚拟情境能对邻里关系的种族和社会成分进行更多不张扬的干预，这样还能减少社会期望偏差。[34]

因此，视频呈现的主要优点是它们降低了调查对象对研究者干预的注意，从而减少了社会期望偏差和实验者需求的影响。此外，视听的呈现可能降低调查对象的认知复杂性，因为调查对象不需要处理长篇的文字性虚拟情境，可以更直接地体验场景。这种方式尤其对那些不习惯回答长问卷的调查对象（例如孩子或教育程度较低、阅读能力低下或参与调查经验较少的人）来说特别有吸引力。但是，所有类型的图片呈现也会受较低标准的实验刺激的影响，减损一些内部效度。为了提高不同视频虚拟情境之间的标准化，我们可能需要训练专业演员尽可能严格按照给定的剧本出演。另一种影响内部效度的风险是，图片或视频虚拟情境中的人们和地区（例如调查对象熟知的特定地区或个人）会引发调查对象的特定关联，这样的关联会影响实验刺激的标准化。[35]一个可能的补救方式是模糊处理图片中呈现的人物面孔（Mutz，2011）。出于这样或那样的原因，就时间和研究资源方面而言，这样的视频虚拟情境就相当昂贵了。此外，视频虚拟情境对调查对象而言也更耗时，我们只能呈现一个或几个案例。由此，我们只能在给定数量的调查对象中收集相对较少

的评价。反之,也只能在有限的几个维度中变化。尽管如此,视频虚拟情境应用的优点超越了它们的局限性。当使用视频虚拟情境时,我们建议进行广泛的预测试以确保它们符合预期目的(例如,不张扬但能被调查对象注意到且理解的刺激因素符合研究者的期望)。

小结

我们可以在文字或表格形式中呈现虚拟情境。文字虚拟情境可能有助于调查对象想象在虚拟情境下的自己或与虚拟情境中描述的人物产生共情。在对调查对象的操控中,文本虚拟情境比较隐蔽,这样有助于降低社会期望偏差。表格式的虚拟情境(和其他强调虚拟情境维度的形式)可能有助于调查对象形成更一贯性的判断。表格形式可能更适合一些决策性任务。表格虚拟情境允许研究者改变虚拟情境维度的排序,这有助于避免维度顺序效应。

虚拟情境的位置可能影响评价。这种影响可能的原因是延滞效应、学习效应和疲劳效应,以及应答删失。通过仅为每个调查对象提供一个虚拟情境可以完全避免这些影响,但这会增加研究成本。当对每个调查对象使用几个虚拟情境时,我们应该随机调整虚拟情境排序来中和参数估计的影响。如果问卷一开始就用极端虚拟情境案例,这可能避免了应答删失,但依然弊大于利。

图片或视听呈现虚拟情境的方式能让研究者对调查对象的操控更隐蔽,同时也适用于那些认知水平较低或阅读技能较差的调查对象。但是这种呈现会因为实验刺激的较低

标准化，影响了内部效度。视频虚拟情境的准备和呈现是非常耗时的。与文本虚拟情境相比，它们的成本效益率更低。另外，对于单个调查对象只能呈现一个或少数视频虚拟情境也是抬升成本的一个原因。

第4节 | 调查模式

虽然已经有一些研究通过电话采访收集过因素调查数据(例如 Emerson et al., 2001; Pager & Quillian, 2005),但是至少对于有多个维度的因素调查来说,我们不推荐这种方法。虚拟情境提供的信息太复杂,以至于只以口头方式呈现,就不容易完全理解和记住,至少可能会出现强烈的顺序效应(例如调查对象更能记住虚拟情境最后的维度,从而对评估结果产生更大的影响;见 Auspurg & Jäackle, 2012)。因此,我们强烈建议因素调查的实施要以自我管理的工具来进行(并且能降低社会期望偏差)。这种方法为调查对象提供了回顾虚拟情境的机会。[36]因此,在所有访问员进行的调查中,调查问卷(或电脑)应该在因素调查模块一开始交给调查对象,访问员也会在调查对象需要时提供帮助。尽管访问员可以帮助澄清问题(按照我们的经验,这几乎没必要),但是大多数调查对象可以独立完成应答任务。公平收入的因素调查的结果(以一般群众为样本)在完全自我管理方式和访问员协助的访问方式中几乎没有差别,唯一的小区别是访问员在场可能会导致调查对象更支持收入公平。这种实验者效应也存在于许多实验室的实验中(对于匿名效应,例如见 Franzen & Pointner, 2012)。请注意,需要访问员澄清的

必要性取决于调查的主题和调查对象样本的情况。

因素调查可以以纸笔(PAPI)或计算机辅助的自我管理的访问方式进行(CAPI 或 CASI)。计算机辅助访问能更灵活地提供信息(例如,可以通过帮助按钮在另一个网站上提供关于单一维度或额外的指导信息)和使用图片或视频材料呈现。此外,在只有计算机支持的调查中可以使用相伴式访问,即虚拟情境内容会根据已调查问题或虚拟情境的回答进行调整(例如,见 Abraham et al.,2010;Li,Chang,& Jasso,2007)。除此以外,再没有其他理由可以说明这个方法比其他方法强了(关于调查实验的不同调查模式的更详细的讨论请见 Champ & Welsh,2006)。

小结

在大多数应用中,如果只是以口头方式呈现虚拟情境,调查对象难以理解和记住虚拟情境提供的复杂信息,由此出现的强排序效应会是一个问题。因此,我们应该让调查对象自己阅读虚拟情境(即使在采用调查员实施的模式中,我们也必须将虚拟情境作为自我完成模式呈现出来,即尽可能让调查对象自己完成阅读和作答)。自我完成模式的优势在于减少社会期望偏差。计算机辅助模式比纸笔问卷更能灵活地向调查对象展示虚拟情境。例如,我们可能会使用视听呈现一些或者全部虚拟情境维度。除此以外,这些模式只在一些特征上与一般的调查研究有点差异。

第5节 | 调查问卷的实施

目前,没有一款专门适用于因素调查模块的软件。为相关方法设计的软件包,例如联合分析和选择性实验,通常不能提供足够的灵活性来适应因素调查的特殊设计(例如还不能适用文字的虚拟情境或者评分)。因此,我们还需要使用统计和办公软件在多步骤中创设虚拟情境文本。我们强烈推荐在问卷中使用邮件合并功能,这样能把出现错误的影响和风险降到最低(例如,将错误的虚拟情境复制到问卷中)。我们只要在问卷中为每个调查对象的虚拟情境置入 n_d 个通配符(wildcards),随后*初始数据库*中的虚拟情境文本会上载到此处。这些通配符还能协助为每个调查对象的虚拟情境随机排序。

下面描述的是一种创设虚拟情境问卷的方法。这是在建立了一个虚拟情境样本(实验设计)和起草虚拟情境文本(见第3章)后进行的。第一,把虚拟情境维度的数字代码转译为虚拟情境文本。使用统计软件包创建包含单一虚拟情境文本语句的字符串变量。

第二,随机安排每个调查对象的虚拟情境,以避免顺序效应。这个过程包括:(1)准备等同于调查对象数量的虚拟情境层块,(2)对每个层块中的虚拟情境进行随机洗牌,(3)给不

同问卷版本标注识别号。所有步骤都可以借助统计软件进行。

第三，为了准备在问卷中实施虚拟情境，数据必须被重整为一个宽格式，单一数据行表示单个问卷版本和属于每个层块的虚拟情境不再储存在列当中，而是储存在行当中（更多关于长和宽的数据格式见第5章）。

第四，我们必须为问卷版本随机安排一个顺序，以确保虚拟情境被随机分配到调查对象。这一步完成后，调查问卷可以按"先来先得"的原则发放，即第一个（第 n 个）登记参与调查的人就接受第一个（第 n 个）问卷版本。[37] 所有这些步骤可以通过简单的数据管理技术来实现，例如在所有数据软件包中提供（在下面的网站中可以找到收入公平例子的 Stata 代码：www.sagepub.com/auspurg_hinz）。图4.3显示了一个设置数据的例子。

识别码（问卷和层块）		第一个虚拟情境变量（所有维度的数字变量与虚拟情境文本的字符串变量）					第二个虚拟情境变量			
id_quest	id_deck	sex_1	age_1		vig_1		sex_2	age_2		vig_2
1	4	1	35		A 35-year-old man w		2	25		A 25-year-ol
2	3	1	30		A 30-year-old man w		1	35		A 35-year-ol
3	6	2	25	…	A 25-year-old woman	…	2	50	…	A 50-year-ol
4	5	2	50		A 50-year-old woman		1	55		A 55-year-ol
5	1	1	40		A 40-year-old man w		2	40		A 40-year-ol
6	2	1	60		A 60-year-old man w		1	25		A 25-year-ol

图4.3 设置数据文件示意图

第五，我们必须在纸笔问卷、在线问卷或计算机辅助问卷中输入实验的设置数据。这一步骤可通过将虚拟情境和应答量表的通配符置入问卷中的正确位置来实现。我们应

该确保调查对象可以在虚拟情境间向前或向后移动，以避免应答删失。

最后，在问卷中把包含设置数据中虚拟情境的文本变量与通配符连接起来。纸质问卷可以使用办公软件提供的邮件合并功能以及类似的程序来实现。计算机辅助调查或在线调查可以使用很多的已有程序。我们可以使用为追踪调查开发的模块来显示已存在的调查对象信息，也可以把虚拟情境文本当作预载变量。图 4.4 提供了这个概念的示意图。

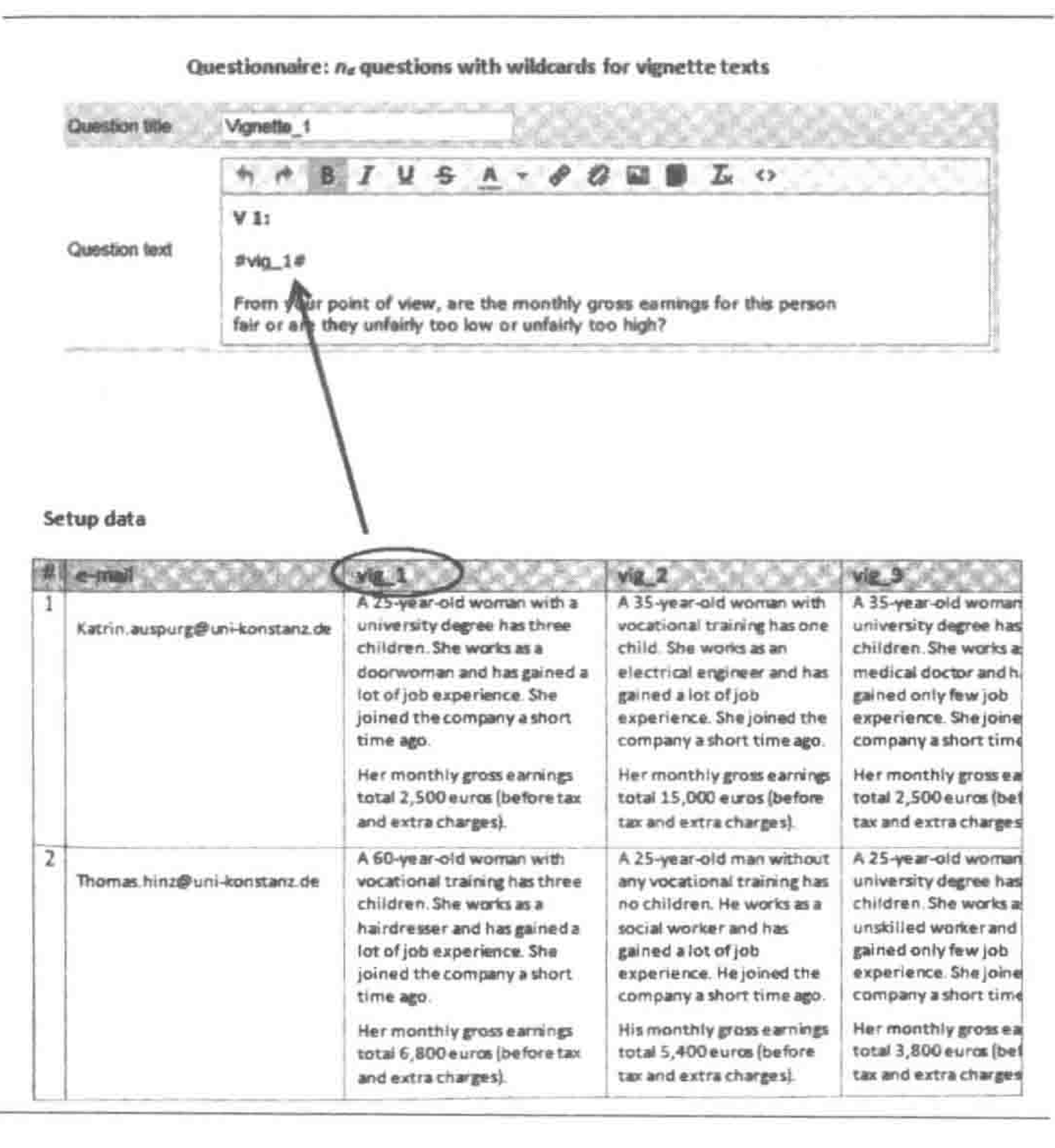


图 4.4　因素调查模块的编程与设置数据相关的通配符

在计算机软件不能使用预载变量的情况下，另一种方法是手动将这些虚拟情境文本复制到问卷中。这种方法可以使用任何软件，但是需要大量额外的工作，并且容易出错，例如出现虚拟情境的混乱。出现这样的错误相当麻烦，因为在数据分析中，研究者很难发现调查对象看到的是无意的虚拟情境。因此，我们建议只使用那些允许预载变量的软件。

若希望突出显示虚拟情境中的某些信息（例如以粗体显示维度水平）或是重组虚拟情境（例如开始每一句话时另起一行），可以把虚拟情境所需的软件代码放在设置数据中（例如 HTML 代码被迫换行）。虽然这些代码只是简单地放在单一的文本短语中，但能减少开发虚拟情境问卷所需的工作量。

无论使用何种技术，我们必须能够识别出哪个调查对象以什么顺序收到哪个虚拟情境。因此，我们必须对调查对象接收到哪个问卷版本有清晰的记录。此外，我们必须确保设置的数据不会丢失，因为所有关于虚拟情境版本的信息和它们对单个调查对象的顺序都储存在这些数据中。在纸质问卷上（例如在页眉或页脚中）打印虚拟情境和问卷版本的标识号（ids）是必要的。如果使用计算机辅助问卷，标识号必须包含在输出的应答数据中。此外，很重要的一点是，对于单个调查对象，虚拟情境是完全随机分配的。和所有实验方法一样，只有在参与者中随机分配实验干预，才能排除其他可能引起参与者反应差异的因素（包括任何不能被研究者观察到的特性）。基于这样或那样的原因，需要多关注虚拟情境在问卷中的具体实施方式以及如何向调查对象分配问卷。

小结

目前还没有专门生成因素调查问卷的软件工具。我们可以通过连接通配符与数据文件的方法将虚拟情境放入问卷中。在这个文件中，虚拟情境文本完全储存在为所有调查对象安排好的顺序中，也包括为生成虚拟情境文本的编码。我们编程的虚拟情境问题必须与调查对象人数相等。如果不使用通配符，我们必须把每一个虚拟情境作为一个单独的问题进行编程。通配符不容易出错，同时便于生成随机的虚拟情境顺序。在纸质问卷中，我们可以通过办公软件提供的邮件合并功能来实现通配符。计算机协助问卷的软件工具也存在类似的选择，允许加载调查对象的预存信息。有必要对不同调查对象呈现的虚拟情境和它们的顺序进行清晰的标识。

第6节 | 给调查对象的指导语

无论选择哪种调查模式，重要的是向调查对象提供一组初始说明，解释任务的性质和应答的量表。这种指导语应该需要达到统一管理调查对象的效果，避免他们的理解存在差异。指导语应该尽可能地简单直接，而且能够标准化调查对象对虚拟情境文本的框架和解释(Louviere, 1988)。尤其是应该使用指导语传达对理解虚拟情境内容重要(比如所有员工都是全职的)的补充信息。在单一虚拟情境中反复提供这些信息会不必要地拉长虚拟情境文本，并可能使调查对象感到厌烦。重要的是确保所有调查对象已经学习并记住了这些信息(这可以通过在指导语中突出显示最重要的句子)。此外，我们应该告诉调查对象，不是所有的虚拟情境例子都必然对应着“真实世界”，因为研究目标可能往往是为了研究那些在真实情况下不会发生或还没发生的情况。这种指导信息能预防调查对象因读到不合情理的虚拟情境而产生的刺激感。指导语告诉调查对象一些维度是随机加入虚拟情境(例如收入的数量)，这也有助于减轻锚定效应(Jasso, 2006, 2012)。最后，指导语应该要求调查对象完全按照自己的信念进行评估，而且他们可以随时修改自己的判断。在杰索(Jasso, 2006)的研究中可以找到有关这方面的指导语示例。

小结

我们应该在因素调查模块一开始时就显示简单且直接的指导语,以提高和标准化调查对象对虚拟情境和回答任务的理解。这些指导语应该包括所有在虚拟情境中保持不变的信息。此外,我们要告知调查对象,一些或全部维度是随机挑选的,并可能有一些超越现实的情况。这种信息可以减轻调查对象的刺激反应,降低可能的锚定效应。

第7节 预测试

从调查对象的角度来看，由于因素调查的应答任务相对于"典型"调查更加独特，要求更多，所以我们应该至少进行一次预测试。在多种维度同时变化时，如果没有预测试，就很难估计现实中调查对象如何看待这些虚拟情境案例的。调查对象可能会认为虚拟情境很假，这就会危及外部效度。此外，他们可能不明白或可能错误地解释一些维度，或者需要一些补充信息帮助他们理解，这都可能会影响内部效度和判断的信度。在调查对象看来，所有虚拟情境可能是极端事例，这就导致了极端估计。相应的评估结果方差值低又将导致较低的统计效力。理想情况下，我们会创设使方差值最大化的虚拟情境案例。这意味着为调查对象成功设计了有意义的评估问题，为识别判断原则提供了较高的统计效力。因此，对因素调查工具进行预测试的一个目的是了解应答分布。需要的话，我们通常可以通过调整一些维度的水平来提高应答的方差（例如，我们可以改变在虚拟情境中呈现的收入范围）。

预测试的另一重动机是检查对虚拟情境和应答任务的理解情况，确定是否有足够的信息来确保对虚拟情境案例的理解是一致的（即高标准化的干预）。此外，我们应该寻找认

知超载的迹象。这种方法论问题可以从调查对象和受访者口头或书面的反馈中进行推断。进一步的迹象包括回归分析或维度中较低的解释性方差，它预测了那些在虚拟情境评估中显示不出来或超出预期的影响。另外，我们可以寻找从已有调查研究中找到的所有其他形式的有问题的回应行为(例如应答集或高水平的无应答单元，在第6章第2节提供了如何识别这些问题的建议)。

当向单个调查对象呈现若干个虚拟情境时，来自一些调查对象的数据足以进行多元分析(例如，有10个调查对象，每个调查对象对10个虚拟情境进行评分，我们就可以收集到100个不同的判断)。强烈推荐进行初步的数据分析，以确保方法的成功。是否可以识别出哪些虚拟情境呈现给哪些调查对象？虚拟情境维度是否不相关且平衡(在这种意义上，单一维度的所有水平都以相同频率发生)？如果不是，那么很可能无法对调查对象随机分配虚拟情境，或者部分设置数据可能意外地没有被使用。

当有明确动机这么做时，调查对象和访问可能会夸大可能的应答问题。根据我们的经验，预测者通常建议在虚拟情境中增加更多的信息。我们不应该对这一建议感到惊讶。好的实验并不需要描绘现实的方方面面，而是显示出某种水平的描述，重点关注最重要的因素(维度)。因此，在大多数情况下，我们建议只增加那些被无数调查对象反复要求的维度，或为了完成虚拟情境任务而明显需要的维度。此外，其他因素可能会作为常数因子被添加到指导语中。

尽管如此，几乎所有的预测试都表明我们应该调整因素调查模块、指导语，或问卷其他部分的某些方面。如果我们

遵循指导方针——如何建构数据设置并将虚拟情境置入问卷中——那么修改问卷的时间和资源是最少的。我们必须简单地做出符合期望的改变(例如对于虚拟情境的措辞),重新运行创建数据设置的统计语法,并将新的数据设置结果与问卷重新连接。因此,我们可以期望在几分钟内修改问卷。相比之下,手工创建并将不同的虚拟情境置入不同的问卷版本这样的任务可能需要几天时间。这就是为什么我们强烈建议在问卷中使用通配符,并使用统计软件包的语法来创建数据设置的原因之一。

小结

值得提醒的是,没有一项田野调查可以在没有预测试的情况下进行,对于因素调查实验尤其如此。我们应该仔细测试调查对象是否理解虚拟情境和应答任务,调查对象和访问员的反馈可以被用于此。此外,分析预测试的数据可以通过高比例的无应答、应答集或其他类型的较低数据质量来发现问题。理想情况下,应答数据的方差应该较高。如果应答在应答量表的特定点上不断积累,那我们可能会改变维度的范围来增强方差。

第5章

数据分析

分析因素调查数据很直接。在大多数情况下，研究者最希望指导虚拟情境变量对测量结果的影响（例如对公平收入的判断）。为调查对象成功地随机化分配虚拟情境，甚至双变量的统计计算，例如比较各组的均值，都提供了有意义的结果。[38]然而，标准的方法是使用多元线性回归模型，所有虚拟情境变量作为自变量。本章的主要目的是考虑在调查对象评估不止一个虚拟情境的因素调查中出现的具体数据结构。目前大多数的典型因素调查都是评估多元虚拟情境。在之前关于统计效力计算的章节中（第 3 章第 5 节第二部分）提到过这个分层数据结构相关的问题。在虚拟情境水平（L1）随机错误 ϵ_{ij} 之外，判断超过一个虚拟情境调查对象（L2）也贡献了附加误差。接下来的数据分析方法主要解决这个问题。在我们讨论统计模型（第 5 章第 2—4 节）之前，我们将关注数据的准备工作（第 5 章第 1 节）。

第 1 节 | 数据的准备

能满足充分分析要求的因素调查数据格式，在外观上与其他分层数据格式相似。数据矩阵中的每一行都包含了虚拟情境变量和结果，调查对象特征随后。调查对象和虚拟情境必须有独特的标识号。如果采用了虚拟情境层块（通常是案例），这些层块也必须有独特的标识。同样地，当采用随机顺序的虚拟情境时，就可以识别出要找的那份的问卷。我们建议在单独一列中记录虚拟情境顺序，以控制可能的顺序效应。

通常，原始数据必须被重新排列以创建合适的格式。原始数据可能被组织起来以便在数据矩阵中为每个调查对象创造一行新的原始数据。在不同的列中，每一个层块或调查对象有 n_d 个结果和虚拟情境标识。通常，这种数据格式被称为宽格式（wide format），而有单一虚拟情境（不是调查对象）呈现不同数据行的分层数据格式被称为长格式（long format）。单独的数据文件提供一个包含虚拟情境变量和标识号的表格（即研究者创设的用于建造单一问卷的数据设置）。研究者可以使用统计软件中的数据管理模式，例如 Stata，来创建如上所述的分层数据格式。然后使用 Stata 中必要的指令 merge 和 reshape。Reshape 指令将宽格式（行中的调查对

象)转变为长格式(列中虚拟情境)。通过使用层块(或问卷)标识和虚拟情境顺序标识作为核心变量,使用 merge 指令,表格内的虚拟情境变量(也应该存在于长格式中)可以被合并到应答的数据中(也重塑了长格式)。当在时间形式的行中或在其他多水平研究的应用中重复测量,例如学生(L1)在教室和/或学校(L2 或更高水平),在因素调查分析中使用的长格式相当于面板数据分析中的数据结构。图 5.1 提供了在长格式和宽格式中的虚拟情境数据示例。任何一个统计包都可以用适当格式准备一个数据文件。

长格式:
每个维度一列;
在应答者中有多少虚拟情境估计就有多少行

id_deck	vignr	sex	age	children	
1	1	female	50	2	→区块 1 的虚拟情境一
1	2	female	40	0	→区块 1 的虚拟情境二
1	3	male	55	2	
1	4	male	50	1	
1	5	female	50	4	
1	6	male	25	0	
1	7	female	55	2	
1	8	male	45	0	
1	9	female	30	0	
1	10	female	40	1	→区块 1 的虚拟情境十
2	1	male	30	1	→区块 2 的虚拟情境一
2	2	male	55	0	
2	3	female	25	0	

宽格式:
每个维度有
n_d 列;
有多少应答者就有多少行

	虚拟情境一				虚拟情境二		
id_deck	sex_1	age_1	children_1		sex_2	age_2	children_2
1	female	50	2		female	40	0
2	male	30	1		male	55	0
3	male	40	0	…	male	60	1
4	male	30	3		female	25	3
5	female	50	0		male	60	3
6	male	40	4		male	30	0
7	male	55	1		male	40	0
8	male	40	4		male	25	3

图 5.1 长数据格式和宽数据格式

在我们开始分析前，必须先做一些检测。第一，研究者应该检查数据矩阵以评估合并和重塑是否成功。第二，要检查虚拟情境变量（以及最终的交互作用术语）和调查对象特征之间的相关性，以确保随机化成功。第三，调查者可以很容易地获得结果的分布频率和方差，以识别出数据结构中可能出现的问题特征，比如对应答数据的较大曲解或应答删失（进一步的细节见第5章第4节）。第四，通过计算调查对象水平上（L2）的方差，可以对应答集进行测试。网站上提供了一个Stata代码的数据准备和初步分析的例子（www.sagepub.com/auspurg_hinz）。在完成详细的检测后，就可以分析数据了。大多数的数据分析要求数据在长格式中，例如下一节讲的回归模型。

小结

分析因素调查数据通常需要预先完成一些数据管理程序。包含调查者对虚拟情境估计的原始数据在随后的列中（宽格式）。它必须被重新安排以便将虚拟情境估计和虚拟情境变量储存在数据矩阵的独立行中（长格式）。调查者必须确保所有的数据集来自不同的来源（虚拟情境数据设置和调查数据），为虚拟情境层块、虚拟情境顺序和调查对象分配必需的识别号。最优的数据结构是带有虚拟情境变量的分层矩阵，包括构成L1特征的因变量和虚拟情境维度，以及构成L2特征的调查对象变量。

第 2 节 | 回归技术

在这本书的前面部分，我们将线性回归模型作为数据分析的标准参照点。最重要的是，必须对结果变量的量表调节进行一般性的假设。本节概述的方法的前提是在一个等距量表中进行测量评估。相关的方法，例如选择性实验，在结果变量的测量上存在差异（见第 6 章第 1 节）。此外，有关评分量表是否达到等距的水平具有争议。根据我们的经验，至少在使用我们推荐的 11 个点的评分量表时，使用类别变量结果的回归方程时结果不会随之改变。由于在这种情况下它们能给出更简单的解释，我们可能倾向于线性模型。这些模型在大多数因素调查应用中都使用（Wallander，2009）。到本节末尾我们会再回到结果变量的测量水平。

在心理学中，相对于回归模型对实验数据的估计，研究者更多使用方差分析（analysis of variance，ANOVA）技术。方差分析的逻辑与线性回归类似，在接下来的部分会重点阐述。

分层数据模型

每个数据分析的总体目标是检测自变量和因变量之间

的系统性相关结构。因素调查有两种类型的自变量。一类是由实验设计所定义的(即虚拟情境变量在它们水平上的变化)。这种分析的目的是检测这些虚拟情境变量和结果的协方差。根据实验设计和分辨度,可以识别虚拟情境变量的交互作用项和它们在结果上的影响。第二类自变量是调查对象的特征。请注意在第 2 章第 1 节呈现的简单回归公式现在有两个方面的修改。首先,增加了调查对象的两个特征变量 Z_1 和 Z_q(例如调查对象的性别和年龄)。其次,模型中添加了误差分量 u_j。虚拟情境变量的影响系数表示为 β(β_1 到 β_p),γ(γ_1 到 γ_q)为调查对象的水平:

$$Y_{ij}=\beta_0+\beta_1 X_{ij1}+\beta_2 X_{ij2}+\cdots+\beta_p X_{ijp}+\gamma_1 Z_{j1}+\gamma_2 Z_{j2}+\cdots+\gamma_q Z_{jq}+u_j+\varepsilon_{ij}$$

$$\text{其中},\ i=1,\ \cdots,\ n_d;\ j=1,\ \cdots,\ n_r \qquad [5.1]$$

这两个新增内容都反映了前面所概述的分层数据结构。因变量 Y_{ij} 的方差可以被分解为两种来源,即调查对象内部的方差和调查对象之间的方差。方差的分解与方差分析技术中使用了相同的策略。统计分析的目的是识别出虚拟情境和调查对象水平上的影响。如前所述,β(β_1, …, β_p)是在虚拟情境水平(L1)上的影响系数,γ(γ_1, …, γ_q)是调查对象水平(L2)的影响。在我们进一步细化这个模型之前(例如跨水平的交互作用和应答删失),我们将更详细地讨论这个简单的模型。

如果研究者只关注虚拟情境变量,这个模版将简化为:

$$Y_{ij}=\beta_0+\beta_1 X_{ij1}+\beta_2 X_{ij2}+\cdots+\beta_p X_{ijp}+u_j+\varepsilon_{ij}$$

$$\text{其中},\ i=1,\ \cdots,\ n_d;\ j=1,\ \cdots,\ n_r \qquad [5.2]$$

这个公式与公式 2.1 之间的差别在于误差分量 u_j[截距和这个误差分量的总和(即 β_0+u_j)常被称为随机截距(random intercept, RI)]。误差项被分解为两部分,因为每个调查对象都评估不止一个虚拟情境。因此,正如第 3 章第 5 节第二部分在计算效力中指出的,这就拒绝了在 L1 上独立观测的假设。如果忽视这个结果[使用研究者熟知的最小二乘法(OLS)],回归系数的估计就是无偏的,但系数的标准误是有偏差的(通常被低估了)。换言之,可能很容易拒绝自变量和结果变量之间没有关系的零假设,这会导致不正确的实质性结论。参照方差分析技术,该模型代表了一种重复测量的方差分析(Vonesh & Chinchilli, 1997)。

为了获得方差分解的实证印象,我们可以计算 Y_{ij} 的组内系数 ρ。这个系数的变化范围是 0 到 1,表示 Y_{ij} 的方差回归到 L2 调查对象水平的比例。换言之,这个系数表明结果的方差是不同调查对象对虚拟情境评估的反映。组内相关越高,使用简单 OLS 估计标准误的偏差就越大。正如我们可以从统计效力计算的部分推断出,当每个调查对象的虚拟情境数量相对高而调查对象的总数量相对低时,这个问题就变得更严峻了。

到目前为止,我们已经讨论了在调查对象中聚合虚拟情境。在数据分析中有两种不同的处理方式。首先,只对虚拟情境变量(L1)的影响感兴趣的调查者可以计算出稳健聚类标准误。对稳健聚类标准误进行估计以调节不平等变量的 L1 误差项 ε_{ij}(方差不齐性, heteroscedasticity),它是由聚类引起的。因此,这些值将聚类定义为一种麻烦的来源,但不会对误差结构进行更详细的建模[在 Stata 回归指令中的选

项“vce(cluster)”；Rogers，1994]。请注意，我们不仅使用稳健标准误，而且使用稳健聚类标准误。其次，研究者使用明确关注误差项的多元结构的多元回归模型。在这些模型中，我们试图从考虑斜率参数(β)和人际间的变量的更复杂的模型中分辨出表达不同(个体)的结果值域(例如，对收入公平的估计)。前者被称为随机截距模型(random intercept models)，而后者被称为随机斜率模型(random slope models)。要强调的是，我们主要关注稳健聚类模型或多元模型策略，因为它们最适合因素调查的数据结构。尤其与另一种评估程序比较，这种策略更有效，例如之前应用于因素调查的两步方法(Hox et al.，1991)。在两步的程序中，一些研究者(如 Jasso，2006)建议每个调查对象单独估计 β 系数 β_k(通过每个调查对象进行单独的回归)，这些斜率而后被用作第二步估计程序的结果。这个策略问题在于“跳跃的 β”问题，即 L1 中的 β 估计有很高的标准误，反映了每一个调查对象必须估计数量相对较少的虚拟情境(Hox et al.，1991；Raudenbush & Bryk，2002)。[39]由于其更高的效率，我们建议对 L1 和 L2 变量系数进行同时估计。

公式 5.1 和公式 5.2 包含没有任何随机斜率的随机截距模型。由于空间有限，我们没有为随机斜率模型呈现公式，而是向感兴趣的读者推荐正在不断兴盛的多元分析文献(Rabe-Hesketh & Skrondal，2008；Raudenbush & Bryk，2002；Snijders & Bosker，2012)。当我们预期估计随机斜率模型时，基本的想法是建构随机变量 β 的模型。换言之，我们在理论上可以相信截距和虚拟情境变量的效应在调查对象间是“随机”分布。我们可以通过包含调查对象特征的内

容来试图解释这种随机变化。这个策略涉及对额外参数的估计。例如，在一个随机斜率模型中估计两个额外参数：调查对象中 β 系数的方差和带有随机截距的协方差。随机斜率模型的参数估计，特别是有若干个随机斜率，可能不会收敛(Rabe-Hesketh & Skrondal, 2008: 172)。更重要的是，我们通常对方差层块的分布提出很强的假设。尤其是虚拟变量的斜率(如，虚拟情境中的人物在不同性别之间的群体差异对于调查对象来说是"随机"变化的)，在估计随机斜率时，假设"妨害"参数遵循正态分布，这可能是值得怀疑的。因为研究者主要调查了虚拟情境变量的影响，以及在一些情况下调查对象间的变化，所以一个更简洁的模型是合适的。因此，我们建议研究者使用随机截距模型，并从只包含虚拟情境变量的模型开始。我们在第二步增加了调查对象水平的变量，在最后一步调查了理论上的跨水平交互作用。在跨水平效应存在的情况下，对跨水平交互作用系数的统计测试的效力相较于虚拟情境变量的随机斜率模型更高(Snijders & Bosker, 2012: 82)。

多水平方法适用于一切因素调查研究策略，尤其当评估不公平时。正如罗西和安德森(Rossi & Anderson, 1982)所主张的，进行因素调查数据分析的目的在于将判断过程中的社会性的部分和特殊的部分剥离开。虚拟情境变量的影响呈现了虚拟情境维度如何影响调查对象判断的一个共享的社会理解，而调查对象变量的影响和跨水平交互作用表明调查对象间子群体的不同看法。然而，在截距和斜率参数中的随机变量反映了判断中的特殊性。

为了控制可能的层块影响，我们可能还会纳入虚拟变量

(Atzmüller & Steiner，2010)。对于这些变量的处理可以类似于其他L2变量。[40]然而，当使用D-efficient抽样技术阻止虚拟情境样本进入不同的层块时，我们通常不希望出现太强烈的层块效应。

随机或固定效应

值得注意的是，上面提到的多水平模型——随机截距和随机效率模型——与固定效应(FE)相比，都是随机效应(RE)模型。固定效应模型直观地包含了特定调查对象的截距(0/1-虚拟变量识别每个单一调查对象)。每个调查对象有不同水平的应答，与随机效应模型相反，这是假定对一个固定参数而不是随机参数进行估计。换言之，L2误差(随机截距u_j)被假定为只从随机效应模型的影响样本中的随机选择。然而，固定效应模型在控制面板研究中不随时间变化的变量的未被观察到的异质性方面具有吸引力(Allison，2009)，它们对因素调查的数据分析不够理想。调查对象变量的影响通常尤为有趣。固定效应方法限制了我们越过调查对象群体对不同模型进行比较(例如男性与女性调查对象的对比)或对跨水平交互作用的比较。[41]这些子群体的比较不存在问题，我们可以从不同的子群体中清楚地测试β系数，从而拒绝群体间不存在差异的零假设(这类似于在综合的分析中分别测试跨水平的交互作用)。然而，为了同时对比几个不同的调查对象特征，我们很快就会面临在过量子群体中案例的数量很少的问题(或使用过量跨水平交互作用时多重共线性的问题)。此外，这些实际考虑之外的因素使随

机效应法在因素调查研究中合理化。假设虚拟情境(刺激)的随机化对于调查对象的随机样本是有效的,因素调查数据实现的首要条件是使用随机效应模型[即协方差与误差项无关(从技术上来讲,corr(X_k, u_j)=0;corr(Z_m, u_j)=0;Cameron & Trivedi,2009:232)]。至少对于L1(虚拟情境)的变量,这个假设是通过设计实现的。因此,借由成功的随机化,可以采用随机效应模型来持续估计虚拟情境维度的影响。结果是,可以从比固定效应模型统计效力更高的随机效应模型中获益(Hausman,1978:1263)。在这种情况下,随机效应模型比固定效应模型更有效,因为随机效应的估计量有更多的方差,产生更低的标准误。

回归分析的示例说明

为了说明数据分析,我们将使用因素调查研究中公平收入的数据。数据是2008年在德国西南部一个小城市(康斯坦茨市)的成人群体调查中收集的。有超过400名调查对象参与调查,调查采用了常见的抽样技术。在因素调查模块中,所有调查对象使用11点评分量表(−5=不公平地低,一直到+5=不公平地高;见图5.2)对公平收入的10个不同的虚拟情境进行评估。雇员的特点有8个维度(包括总收入),这是基于先前研究已知的会影响公平评估的内容(Jasso & Webster,1997,1999)。第3章的表3.1已经描述了虚拟情境变量:性别(1=女性,0=男性),年龄(从25岁至60岁,五年为一个区间),受教育程度(职业培训或大学学位,未受训练是参照类别),生育的孩子数量(0、1、2或3),职业声望

[10类不同职业的等级声望分数(MPSs),从没有技术的工人到医生],工作经验(1=经验丰富,0=经验不足),工作年限(1=长,2=短)。我们生成了一个 *D*-efficient 样本,有24个不同层块,每组10个虚拟情境,排除了不合逻辑的情况(例如,医生未取得大学学位)。我们使用了分辨度IV的设计,所有的主效应和一些社会理论及先前研究中所期待的双向交互作用(例如,在性别和劳动力市场特征之间的交互作用)被正交化。所使用的虚拟情境样本的 *D*-efficiency 达到90.4。层块被随机分配给调查对象。所有数据是通过纸笔访问收集的。在一些初始问题之后,采用因素调查模块作为一个自我管理模块。拥有大学学位的调查对象被划分为具有较高的受教育水平的组(36.3%)。我们在下面示例中使用的虚拟情境评估的数量为 n=2 474。为了简单起见,我们避免了模型中虚拟情境变量间的双向交互性,但是在实验设置中考虑到了这些交互性。

一名50岁的未经职业培训的女人,她有两个孩子。她是一名职员,有着丰富的工作经验。

她在公司工作了很长时间。她的月收入总额为1 200欧元(税前且未扣除额外费用)。

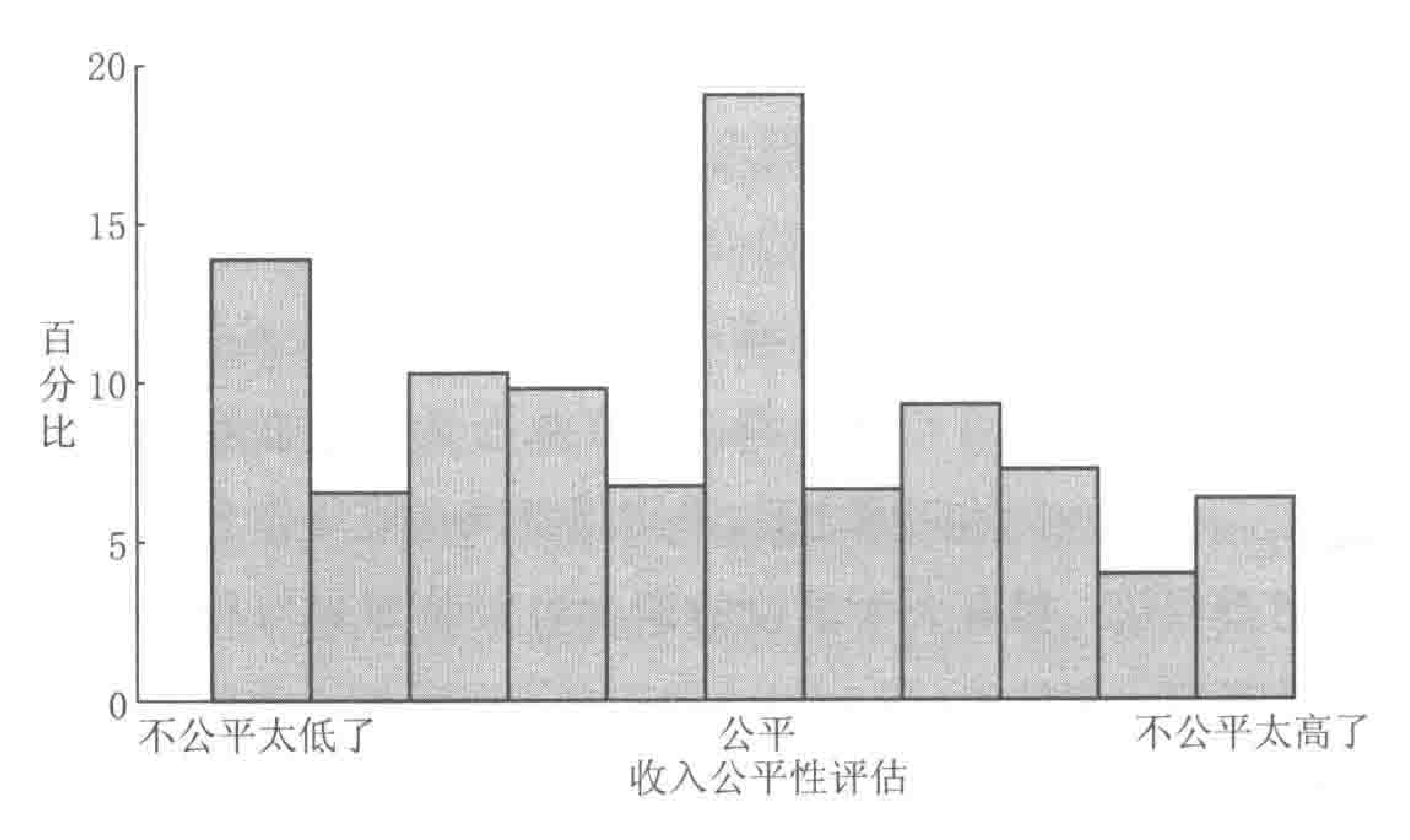
在你看来,这个人的月收入公平吗?是太高还是太低?

不公平 太低了					公平					不公平 太高了
−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5
■	■	■	■	■	■	■	■	■	■	■

资料来源:Justice of Earnings Survey Konstanz,2008。

图5.2 收入公平的因素调查虚拟情境例子(变异的维度用粗体表示)

虚拟情境评估分布的频次显示了调查对象在量表中使用的所有值(见图 5.3)。但是可以观察到两个峰值出现在应答值—5(不公平地低)和 0(公平)上。当我们讨论应答删失的另一种模型时,我们将参考这个特定的分配模式。



资料来源:Justice of Earnings Survey Konstanz, 2008。

图 5.3 虚拟情境评估频次分布

表 5.1 的第一列显示了 OLS 模型的参数,不考虑更复杂的误差结构。效果系数是直观的。例如,β 系数为 0.265 的 *vig*: *sex_f* 表明虚拟情境中对女性收入的估计比男性高 0.265 点(在评估量表中)。换言之,女性更可能被评估为报酬过高。第二列包含了与聚类标准误的评估。从第一列到第二列,虚拟情境变量的标准误发生了变化,对于一些变量而言,它们稍微增长了,对于另一些变量则收缩了。该结果是由于在集群内的异质性的不同模式所引起的,它只能被考虑进第二列。在我们的例子中,变化并不显著。所有系数在两个模型中都具有较高的显著性。[42]但是,我们应该注意嵌套的数据结构,尤其是如果分析了调查对象的更小的子群体。在下

一步中，要估计虚拟情境变量的随机截距模型（见表 5.1 的第三列）。u_j 的方差参数表明了在调查对象的值域（截距）中有显著性变化。这个结果可以从截距的标准偏差（u_j）中观察到，值为 0.617。我们可以很容易地计算出当前例子中的组内相关系数 ρ 为 0.116（自动提供 Stata 命令 *xtreg*）。[43]这是一个相对适中的值（即，在判断中只有 12％的变量归因为调查对象间的变化）。换言之，对公平的评估在调查对象间是一致的。一个似然比（likelihood ratio，LR）测试（测试值是 χ^2 分布）适用于所有嵌套模型的比较，随机截距模型比忽视在简单的 OLS 模型中的分层结构更明显地适合数据（χ^2 值：100.47；df＝1；$p<0.000\,1$）。这样一个 LR 测试的结果是在采用 Stata 命令 *xtreg* 后自动呈现的（*xtreg* 命令是选项“fe”），我们也接收一个样本测试，即 RE 模型的假设的因变量是否与 L2 误差项 u_j 无关。[44]类似的测试可以用其他统计包执行。

表 5.1　公平评估的回归模型（OLS，带有聚类稳健标准误的 OLS，以及随机截距模型）

	（1）OLS	（2）OLS 和聚类稳健标准误	（3）随机截距
Vig：*sex_f*	0.265***	0.265***	0.259***
	(0.073 6)	(0.085 4)	(0.081 5)
Vig：*age*	−0.010 5***	−0.010 5***	−0.010 9***
	(0.003 20)	(0.003 06)	(0.003 03)
Vig：*vocational training*	−0.394***	−0.394***	−0.386***
	(0.091 1)	(0.080 6)	(0.086 0)
Vig：*university*	−0.462***	−0.462***	−0.444***
	(0.092 6)	(0.084 9)	(0.088 1)
Vig：*#children*	−0.141***	−0.141***	−0.141***
	(0.024 6)	(0.027 3)	(0.023 3)
Vig：*prestige*	−0.014 1***	−0.014 1***	−0.014 4***
	(0.000 995)	(0.000 930)	(0.000 942)

续表

	(1) OLS	(2) OLS 和聚类稳健标准误	(3) 随机截距
Vig: *experience*(*long*)	−0.422***	−0.422***	−0.423***
	(0.097 1)	(0.100)	(0.092 0)
Vig: *tenure*(*long*)	−0.219**	−0.219**	−0.210**
	(0.089 3)	(0.084 7)	(0.084 6)
Vig: *log_income*	2.521***	2.521***	2.532***
	(0.041 8)	(0.060 1)	(0.039 5)
常数	−18.20***	−18.20***	−18.24***
	(0.338)	(0.441)	(0.324)
虚拟情境数量	2 474	2 474	2 474
调查对象数量	249	249	249
R^2	0.629	0.629	
对数似然函数值	−4 986.13	−4 986.13	−4 935.90
标准偏差 u_j			0.617
标准偏差 ε_{ij}			1.708
组内相关系数 ρ			0.116

注:括号中是标准误。

** $p<0.05$, *** $p<0.01$。

资料来源:Justice of Earnings Survey Konstanz, 2008。

下一步是确定调查对象特征如何影响结果。特别是调查对象水平(L2)变量的模型必须适用于分层数据结构。从演示的目的而言,我们只关注一个调查对象水平的变量——受教育程度高(*educ_high*)——调查对象是否获得大学学位的指标(0/1)。正如表 5.2 第一列所示,*educ_high* 的系数是 0.231($p<0.05$)(即,*educ_high* 作为主效应的模型)。如果我们在 OLS 模型中包含这个变量但不考虑嵌套结构,系数会错误地表现出高显著性。那么,应该如何解释系数呢?受教育水平较高的调查对象对公平收入的评分相比参照组(没有获得大学学位)的调查对象高了平均 0.231 点(在评估量表中)。似然化测试表明,包含调查对象水平的变量可以提高

数据的适配性。表5.2中的第二、第三列包含了这两组的独立模型系数。有趣的是，没有大学学位的调查对象认为受教育程度与收入的公平性更相关。例如，与 *educ_high* 组相比，受教育程度较低的组的虚拟情境变量中，职业培训的绝对值是前者的一倍以上（−0.218 相比 −0.485）。在虚拟情境变量中，大学学位上也呈现类似趋势（−0.313 相比 −0.519）。有两种不同的方式来测试第二列与第三列模型之间的差异是否具有显著性。首先，可以使用 F 检验（通常也称“Chow-Test”；见 Wooldridge，2013），考虑到同时影响判断的所有虚拟情境变量。必须建立与每个虚拟情境变量与 *educ_high* 的交互作用项。然后，一个为所有调查对象建立的“集合性”模型包含所有虚拟情境变量、*educ_high* 以及所有交互作用项。F 测试的目的是拒绝零假设，即所有交互作用项都同时为零。在我们的例子中，F 测试并没有拒绝受教育程度较高和较低的调查对象之间平等评估原则的零假设[F(9 248=1.10；p=0.364 0)]。其次，我们可以只建立和测试特定兴趣变量的跨水平交互作用项，对这些项的显著性水平使用标准 t 检验。我们的例子关注调查对象的教育和虚拟情境变量的教育之间的跨水平交互性。在表5.2的第四列中呈现了一个带有两个其他交互性影响的随机截距模型。显而易见，调查对象的主效应已经不再显著。这两个跨水平的交互作用都有预期的（正向）信号[即对于拥有大学学位的调查对象，虚拟情境人物的教育受到的（负向）影响减少了]。这两个项都没有达到统计学意义上的显著性。也就是说，我们发现了一个弱交互作用的证据，但是在两个子群体中的案例数量（或统计检验力）是不足以拒绝没有显著的交互作用的零假设的。

表 5.2 公平性评估的回归模型(具有 L2 变量、小组比较和跨水平交互作用的随机截距模型)

	(1) L2 教育程度的随机截距	(2) 教育程度低的随机截距	(3) 教育程度高的随机截距	(4)跨水平交互作用随机截距
Vig: *sex_f*	0.247 ***	0.221 **	0.280 **	0.249 ***
	(0.081 5)	(0.102)	(0.136)	(0.081 6)
Vig: *age*	−0.010 9 ***	−0.008 31 ***	−0.015 2 ***	−0.010 8 ***
	(0.003 02)	(0.003 66)	(0.005 35)	(0.003 02)
Vig: *vocational training*	−0.386 ***	−0.485 ***	−0.218	−0.490 ***
	(0.086 0)	(0.104)	(0.151)	(0.108)
Vig: *university*	−0.443 ***	−0.519 ***	−0.313 **	−0.523 ***
	(0.088 1)	(0.106)	(0.157)	(0.109)
Vig: #*children*	−0.141 ***	−0.150 ***	−0.131 ***	−0.142 ***
	(0.023 3)	(0.028 2)	(0.041 1)	(0.023 3)
Vig: *prestige*	−0.014 4 ***	−0.014 5 ***	−0.014 4 ***	−0.014 4 ***
	(0.000 942)	(0.001 14)	(0.001 66)	(0.000 942)
Vig: *experience*(*long*)	−0.424 ***	−0.429 ***	−0.423 ***	−0.428 ***
	(0.092 0)	(0.111)	(0.162)	(0.092 0)
Vig: *tenure*(*long*)	−0.208 **	−0.198 *	−0.231	−0.210 **
	(0.084 6)	(0.103)	(0.148)	(0.084 6)
Vig: *log_income*	2.533 ***	2.535 ***	2.533 ***	2.534 ***
	(0.039 5)	(0.048 3)	(0.068 5)	(0.039 4)
Resp: *educ_high*	0.231 **			0.056 4
	(0.108)			(0.153)
Vig: *voc trainting* × *resp*: *educ_high* 的跨水平交互作用				0.282
				(0.178)
Vig: *voc trainting* × *resp*: *educ_high* 的跨水平交互作用				0.222
				(0.179)
常数	−18.32 ***	−18.36 ***	−18.05 ***	−18.26 ***
	(0.326)	(0.392)	(0.572)	(0.328)
虚拟情境数量	2 474	1 570	904	2 474
调查对象数量	249	158	91	249
对数似然函数值	−4 933.627	−3 079.672	−1 846.500	−4 932.248
标准偏差 u_j	0.613	0.654	0.513	0.617
标准偏差 ε_{ij}	1.708	1.645	1.818	1.708
组内相关系数 ρ	0.114	0.136	0.074	0.115

注:括号中是标准误。阴影部分强调了调查对象教育程度的影响、虚拟情境人物教育程度的影响,以及他们的交互作用。

* $p<0.1$, ** $p<0.05$, *** $p<0.01$。

资料来源:Justice of Earnings Survey Konstanz, 2008。

确定虚拟情境变量在调查对象间影响的变化程度的一个探索性策略是，对每个虚拟情境变量进行个别且连续的随机斜率模型估计，同时确定一些虚拟情境变量的影响是否有显著的变化。对于当前的数据集，这种探索将导致在职业培训、育儿数量、经验和 *log_income* 的斜率中出现显著的随机变化。但是，研究者应该对调查对象为什么对虚拟情境变量有不同评估有理论上的假设。理想的情况是，应该包括调查对象的解释性特征。

小结

因素调查数据分析必须考虑到因变量的测量类型(即定类、定序，或定比)。使用逻辑回归(见第 6 章第 1 节)对定类变量的结果进行分析。如果我们认为结果是定序变量，那么就需要定序分类模型。大部分因素调查是测量态度或评估的定比变量，这是连续结果使用回归模型的先决条件，如 OLS。然而，嵌套在调查对象内的虚拟情境变量的数据结构需要计算调整(聚类)标准误的技术。否则调查对象变量(L2)的影响通常被错误地估计为显著的。

我们推荐使用聚类稳健标准误的回归模型或多水平模型(引层线性模型，hierarchical linear models，HLM)。随机截距模型假设评估的值域对于调查对象是不同的(随机)。但是，在结果上虚拟情境变量的影响是相同的，因此对评估的社会共享部分进行建模。随机截距模型估算了大量的调查对象间差异。我们建议使用特定子群体随机截距模型或跨水平交互作用来确定在不同调查对象中评估规则的差异。这种方法比带有虚拟情境变量的“随机”变化影响的随机斜率模型更有效。

第 3 节 | 相对效应值和交叉弹性

我们现在强调因素调查数据分析中一个特别重要且具有吸引力的特征。由于多维度设计，可以估计虚拟情境变量的相对效应值以及它们之间的权衡。如果对所有虚拟情境维度使用相同数量的水平，那么所有维度的回归系数的强度具有相互的比较性（请记住，分类的虚拟情境变量可能呈现不同基本单元，这限制了它们效应值的可比性）。否则，就可以计算半偏（semi-partial）R^2 值（例如，使用用户编写的 Stata *pcorr2* 或 *domin*）以测量解释调查对象评估的虚拟情境变量中的不同相关性（这些值测量了方差 R^2 的比例，它是由单一变量解释的）。然而，这些计算中的一些是正确的，只有当被分数化时，才对所有维度采用具有相似相关和方差的 *D*-efficient 设计。[45]此外，我们很容易了解不同虚拟情境变量之间的权衡。这里指的是收入公平性的例子。在上面的分析中，女性的收入相比相同特征的男性而言，被评估为太高了。这个问题是当用欧元这种货币来衡量时，一个合理的性别收入的差距（JGPG）是多少。在公式 5.3 中，X_1 代表了欧元收入的自然对数（如虚拟情境中所示），X_2 代表了虚拟情境中任务的性别（1＝女性，0＝男性）：

$$Y_{ij}=\beta_0+\beta_1\log(\text{收入})_{ij}+\beta_2\text{女性}_{ij}+\cdots+\beta_p X_{ijp}+u_j+\varepsilon_{ij}$$

$$\text{其中}, i=1, \cdots, n_d; j=1, \cdots, n_r \quad [5.3]$$

JGPG 弥补了虚拟情境中人物性别的影响[从技术上讲：$\beta_1\log(\text{收入}+\text{JGPG})_{ij}+\beta_2\text{女性}_{ij}=\beta_3\log(\text{收入})$]。经过一些数值变换后，%JGPG 的数量表示在平均收入上的增加或减少的百分比，计算为 $\%\text{JGPG}=\exp[(-\beta_2/\beta_1)-1]\cdot 100$，这里 exp(·)表示指数函数(Auspurg & Jäckle, 2012)。在包括聚类稳健标准误而没有调查对象协方差的模型中，%JGPG 为 9.97%(精确值为平均每月 480 欧元)。因此，在调查对象眼中，女性的收入大约比男性少 10%才被认为是公平的。同样地，在因素调查中可以计算其他权衡。很明显的是，一个虚拟情境维度必须用连续单位来表示(例如，金钱或者监禁时间)。在货币单位的情况下，这样的权衡等同于支付意愿(willingness to pay, WTP)，用于选择性实验中(见第6章第1节)。对于这个概念一个更常见的术语是交叉弹性(cross-elasticity)。在没有对数转换的情况下，WTP 的公式呈现了一个度量单位或二进制的虚拟情境变量 X_k(编码为1/10)到虚拟情境变量 X_m，在货币单位中呈现，例如欧元或美元，被简单地计算为：

$$\text{WTP}=-\beta_k/\beta_m \quad [5.4]$$

值得注意的是，这些评估依赖于正确的特定模版，并且特别重要的是，所有相关变量都包括在其中。研究人员使用 Stata，可以采用用户编写 *wtp*，在货币单位中提供了 95%的置信区间的交叉弹性(或当使用对数的货币变量时的百分比效应，对于公式和其他在如何估计 WTP 测量的标准误的信

息，见 Hole，2007）。如果在货币单位中的变量是对数，那么 wtp 的结果与正确的估计稍有偏差。

小结

通过因素调查的数据分析，可以判断虚拟情境变量对调查对象判断的强烈程度。如果水平的数量等于所有虚拟情境变量的数量，那么不同变量的回归系数可以直接作为效应量进行比较。否则，每个维度贡献的大小可以通过计算半偏 R^2 值或相关权重的相关测量来表示。因素调查数据的一个具体特征包含（至少）一个自变量是虚拟情境变量间权衡作为交叉弹性进行计算的可能性。在选择性实验中，也可以使用同等的技术来计算 WTP。

第4节 | 应答删失

对虚拟情境评估使用固定回答量表会导致“删失”观察。当在量表中估计一些虚拟情境极端值时(不公平地低或不公平地高),这个问题就会出现。如果可能的话,至少一些调查对象会更极端地评估虚拟情境。对结果变量的分布频率进行检查,可以看到量表中一个或两个端点值显著提高。图5.3显示除了量表的中心点(公平),左边量表终点的评价(不公平地低)与大多数判断选项相比更经常被选择。这些值被限制在量表区间内,这可能会阻止调查对象在极端情况下表达他们的实际评估。在正态的OLS回归中,这种情况会导致回归系数的下降。鉴于这样的情况,计量经济学教科书提出了Tobit分析(Greene, 2003: 764; Tobit, 1958; Wooldridge, 2010: 667 et seq.; 2013: 572 et seq.)。Tobit模型假设了未被观察到的潜在变量(代表在应答量表中的真实分数)、可完全观察到的自变量,以及可观察的结果变量(即在量表中测量评估)。当进行Tobit分析时,我们必须确定误差项的正态分布的假设是否正确。最终,结果变量必须被转变。根据数据结构,更复杂的两阶段过程可能会产生更好的适配。Tobit分析中的这些变量(如, craggit模板: Burke, 2009; Cragg, 1971)试图明确删失过程的建模。我们对Tobit

表 5.3 公平估计的回归模型(随机截距和随机截距 Tobit 模型，具有较低的上限和下限)

	(1) 随机截距	(2) 随机截距 Tobit 模型 下限	(3) 随机截距 Tobit 模型 下限和上限
Vig：*sex_f*	0.259***	0.273***	0.309***
	(0.081 5)	(0.092 9)	(0.099 7)
Vig：*age*	−0.010 9***	−0.013 8***	−0.014 3***
	(0.003 03)	(0.003 41)	(0.003 68)
Vig：*vocational training*	−0.386***	−0.463***	−0.499***
	(0.086 0)	(0.097 0)	(0.105)
Vig：*university*	−0.444***	−0.505***	−0.556***
	(0.088 1)	(0.099 4)	(0.107)
Vig：#*children*	−0.141***	−0.175***	−0.186***
	(0.023 3)	(0.026 4)	(0.028 4)
Vig：*experience*(*long*)	−0.423***	−0.541***	−0.572***
	(0.092 0)	(0.104)	(0.112)
Vig：*tenure*(*long*)	−0.210**	−0.211**	−0.231**
	(0.084 6)	(0.096 0)	(0.103)
Vig：*prestige*	−0.014 1***	−0.015 9***	−0.016 3***
	(0.000 942)	(0.001 06)	(0.001 14)
Vig：*log_income*	2.532***	2.817***	2.927***
	(0.039 5)	(0.046 2)	(0.050 2)
常数	−18.24***	−20.26***	−20.98***
	(0.324)	(0.375)	(0.406)
虚拟情境数量	2 474	2 474	2 474
调查对象数量	249	249	249
左边删失案例		344	344
右边删失案例			157
标准偏差 u_j	0.617	0.735	0.779
标准偏差 ε_{ij}	1.708	1.884	2.014
组内相关系数 ρ	0.116	0.132	0.130

注：括号中是标准误。

* $p<0.1$，** $p<0.05$，*** $p<0.01$。

资料来源：Justice of Earnings Survey Konstanz，2008。

分析的结果进行说明,假设评估量表的下限(－5)或同时假定上限和下限(－5 和＋5;见表 5.3,第二和第三列)。估计策略包括随机截距。我们必须记住 Tobit 模型的系数指的是未被观察到的潜变量(Wooldridge, 2013: 574)。在应答删失的情况下,我们可以认为这种对未被观察到的潜在结果的影响是首要兴趣。[46]

我们通常将随机截距回归模型中的系数与随机截距 Tobit 模型中的系数进行比较(表 5.2 的第一列到第二、第三列)。考虑左边删失和两者限制的影响在某种程度高于(在绝对值上)没有考虑删失量表的模板,尽管这种差距非常小。如果我们有足够的理由相信限制的量表引起了删失结果变量值,那么 Tobit 模型提供了一个更充分的评估模型。但只有当模型重要的假设(如误差项的正态分布或方差齐性)被提前确定时,Tobit 模型才是合理的。请注意,在评估概率时, Tobit 模型有一些扩展,这种概率在分析中包含了调查对象评估虚拟情境的极端值。这种方法在对删失进行建模时更加灵活,我们向感兴趣的读者推荐伍尔德里奇(Wooldridge, 2010)的研究。

我们接下来介绍数据分析中因变量的测量(即调查对象的评估)。如果考虑到要以定序量表测量结果,那么我们应该采用不同的回归模型,例如,有序概率或有序 logit 回归。在我们的案例数据中,我们发现有序 logit 模型的系数与 OLS 系数非常接近。从我们的经验看,尤其是使用 11 点评分量表并且每个调查对象使用 10 个虚拟情境时,无论使用哪种统计模型分析数据,结果都是稳健的。

小结

使用(有限的)评分量表可能会导致应答删失。相对频繁地选择量表的末端值就表示存在删失。如果结果的分布表明删失的测量,可以考虑使用删失回归模型,例如 Tobit。为了处理删失或非测量结果,研究者可以使用有序的 logit 模型或概率模型。

第6章

延展深入

第 1 节 | 相关方法

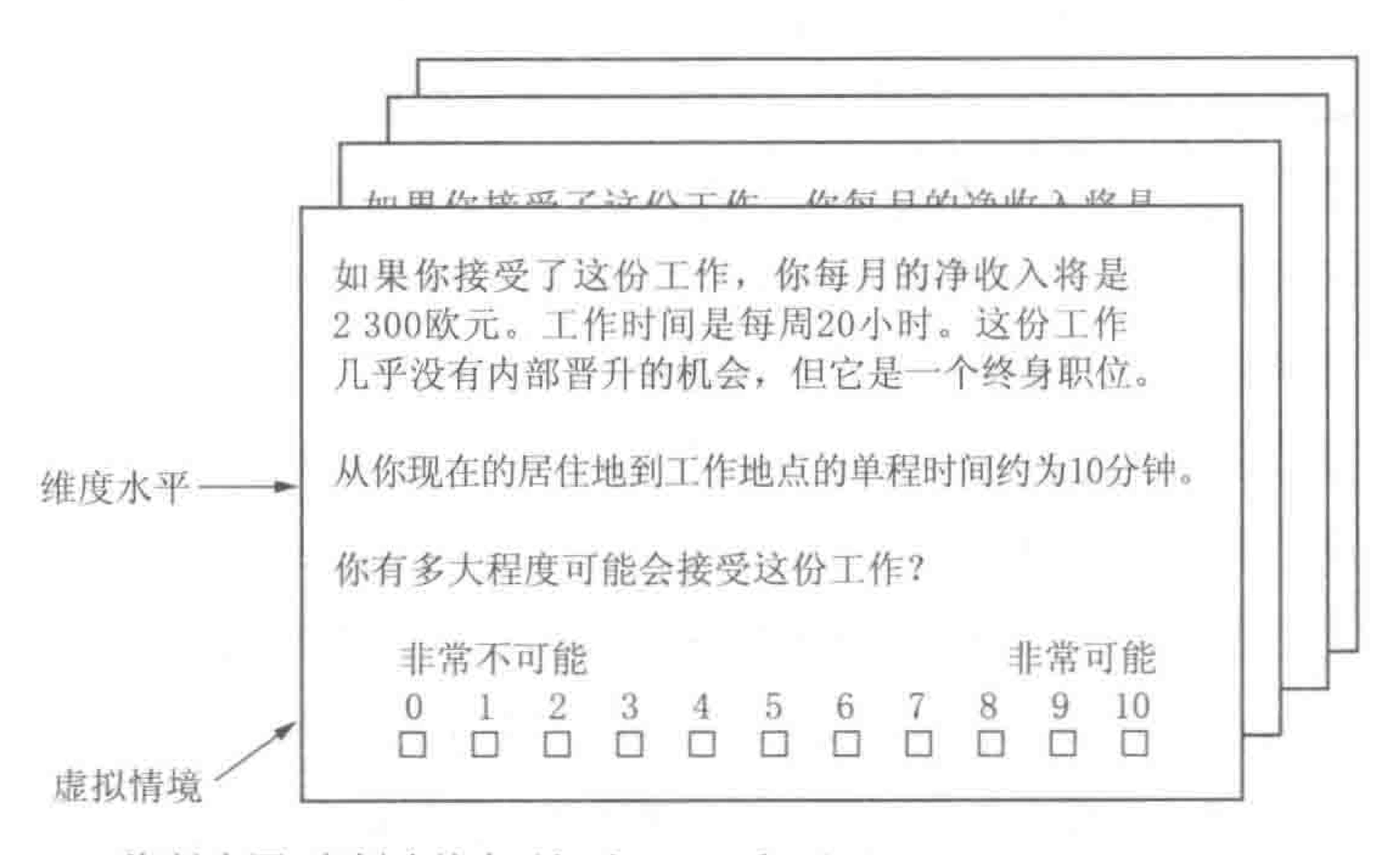
除了基于多因子实验规划的多因素调查外，还存在其他调查方法。其中最突出的方法是联合分析法(conjoint analysis, CJA)和选择性实验(choice experiments, CEs)。这些方法与因素调查存在几个共同点。一些研究者已将联合分析法这个术语作为一个像伞一样的概括性术语覆盖了所有的分解方法，其中对对象和场景的整体评估都被分解为不同维度的影响。然而，在因素调查、选择性实验和联合分析法之间也存在显著差异。这些方法是在不同的研究传统和学术领域中发展起来的。因此，它们在典型设计和数据获得的质量上要求有所不同。因此，区分这些方法是很有帮助的。

联合分析法几乎只被在市场营销研究中使用。大多数的联合分析法是关于消费者和产品或金融和其他服务业的。这样的研究用来评估新产品，分析市场竞争、市场划分以及价格或重新定位对消费者行为的影响(Green & Srinivasan, 1990)。联合分析法的主要目标是衡量不同产品或服务的效用，以及不同特性对于整体效用的贡献。调查对象通常会面对“资料卡”(profile card)上对产品或服务的表格式描述，这些特征(属性)在实验中处于不同水平。因为联合分析法几乎是被商业机构而非学术研究所采用，同时因为它们的动机

是纯粹数学性的(Luce & Tukey, 1964),没有源于行为的社会理论,因此属性和水平通常是临时安排的(Louviere, Flynn, & Carson, 2010)。至于为什么联合分析法没有达到学术社会研究的高标准,还有其他原因。例如,数据分析用的是我们反对的两步法,而不是更有效的多水平分析(见第5章)。虽然多数研究采用了正交主效应的实验设计,但不对交互作用进行估计。调查对象的评分如果一致性不高,或表现出非典型性应答风格的话(即与其他调查对象估计的主要部分不一致的应答),他们的数据常常是被简单剔除的。而在社会科学的学术调查中,这些调查对象却得到了更多的关注。此外,联合分析法通常是根据单个调查对象的偏好来设置属性和特征的。调查对象首先被要求对项目问题的单一属性进行偏好的评分;然后根据他们对属性评分的重要性程度获得相应的资料卡(这称为适应性联合分析法)。这种方法总体上能使用更多不同的属性;但是,实验设计的一个核心特征——向调查对象随机分配刺激——就被损害了。此外,对偏好的直接探问容易产生社会期望偏差,从而消除了调查实验的其他主要优势。[47]总之,联合分析法的实际分辨度可能对他们发展商业研究有效(Green & Srinivasan, 1990; Gustafsson, Herrmann, & Huber, 2007);然而,出于同样的原因,它们不太适合因素调查研究的常规目标。

由于这些原因,我们接下来将选择性实验调查方法的讨论限制在因素调查和选择性实验两者的对比上。这两种方法都是社会科学学术研究具有前景的工具。我们首先重复以下因素调查方法的关键特性。调查对象对若干个对象或情况(虚拟情境)的简短描述进行评分,这些描述由不同的维

度组成，这些维度在实验中处于不同水平。通常情况下，这些虚拟情境被描述为运行的文本，调查对象必须在评分量表上单独对几个虚拟情境进行评估。经典的因素调查处理的是规范的规则和态度或定义，但是因素调查也越来越多地用于调查受访者在虚拟情境描述的情况下可能产生的行为。图 6.1 提供了调查对象对不同工作机会进行评分的应用程序示意图(对这样的应用，见 Abraham et al.，2014；Abraham et al.，2010)。

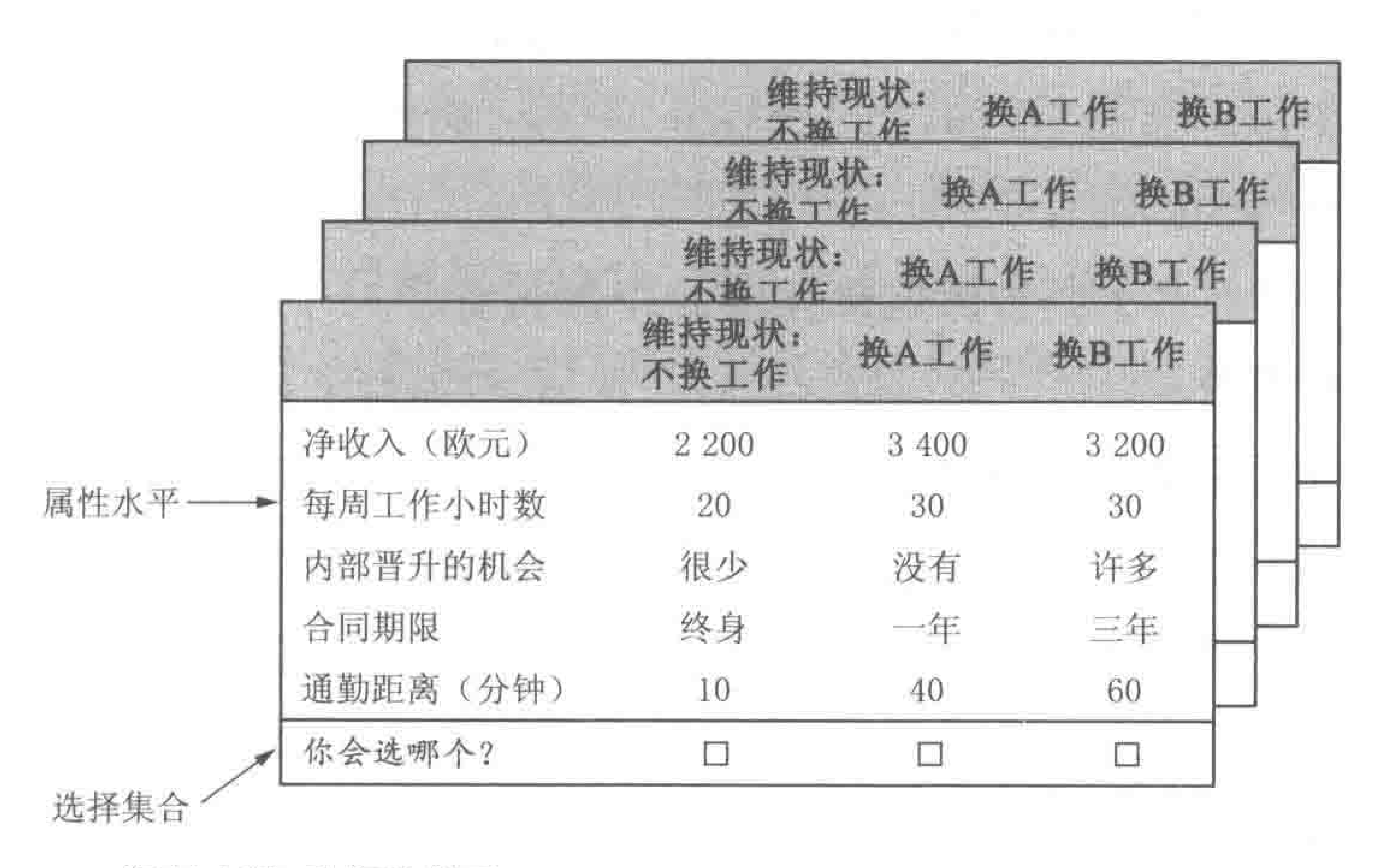


资料来源：案例改编自 Abraham et al.，2010。

图 6.1 因素调查任务：对段落描述(虚拟情境)单独评分

相比之下，选择性实验是一种广泛用于交通研究、环境经济学和健康经济学的方法，用来估计不同选择和它们属性的效用。选择性实验经常用在没有价格信息(或只是补助价格信息)时，了解调查对象对还未存在的选项的偏好估计，或对未面世的产品的支付意愿(WTP)，例如对许多公共服务(如交通、健康服务)和诸如新鲜的空气和干净的水这样的公

共产品的评估。[48]选择性实验典型的应答如图 6.2 所示，图中使用了一个工作机会的例子。调查对象同时面对几个(二、三、四)选项的表格式描述，这些选项在一个单独的选项集合中联合呈现。所有的选项都是由相同的属性组成的，这些属性处于不同的实验水平上。调查对象被问及他们喜欢哪种选择，并面临若干个选择集合(通常是 8 或 10 个)。



资料来源：我们的例子。

图 6.2　选择性实验任务：从一个选择集合中选出一个表格选项

通常情况下，也存在一个不选择或维持现状(status quo)的选项。调查对象可以声明他们不喜欢任何的选项，只保持他们当下的状态。我们建议要提供这一选项，以避免胁迫调查对象做出选择的情况；这样也能更好地评估调查对象的偏好和可能的行为——他们是否真的会使用选项集合中提供的选项或服务？将现状作为参照点也可以对选择选项的绝对效用进行更好的评估。如果没有“不选择”的选项，最常被选中的选项可能不太吸引人，但研究者不可能检测到其低效

用(Louviere et al., 2010)。但缺点就是经常提到的现状偏见(status quo bias),即调查对象比想象的要更多选择此选项(例如,这种选项比真实市场情况下更经常被选择,但它也是一种有效的测量;例如,见 Adamowicz, Boxall, Williams, & Louviere, 1998; Samuelson & Zeckhauser, 1988)。对于采用“不选择”(这也可以被包括于因素调查研究中)感兴趣的读者可以回顾选择性实验的特定文献,这是方法论研究的一个活跃领域(例如,见 Louviere et al., 2010; Meyerhoff & Liebe, 2009)。[49]

在几乎所有的选择性实验中都包含了一个经济属性。应答数据允许对金钱的权衡进行评估,例如对于偏好属性的支付意愿,或对不喜欢属性的避免意愿。为了防止社会期望偏差以及为了创造真实情境,成本通常被设计在更微妙的支付工具中,例如,如果实现了单项选择,调查对象会经历税收百分比的增加。显然,除了金钱之外的测量也可以用来获得相对重要性的说明。例如,选择性实验将等待时间作为支付意愿的替代性变量(例如,见 Ryan, McIntosh, Dean, & Old, 2000)。在第 5 章,我们提供了关于如何获得这些测量(以及相关权衡,例如工资差距)的信息,这些信息可以使用包括了维度的因素调查研究来估计。在因素调查研究中很少使用支付意愿的测量,可能只是因为很多使用者没有意识到这些可能性。

因素调查并没有在一个特定的理论中证立其测量方法(尽管使用者可能会参考罗西的经典文本,在该经典文本中,这个想法首次被介绍:Rossi, 1979; Rossi & Anderson, 1982),但是选择性实验则是由选择行为理论和随机效用理论

(random utility theory，RUT)所驱动的。瑟斯通(Thurstone，1927)首次介绍了随机效用理论，他为使用成对比较来测量效用函数提供了理论依据。随后，麦克法登(McFadden，1976，1986)将随机效用理论扩展到多元比较中(Louviere et al.，2010)。随机效用理论的主要思想植根于效用最大化的一般经济学理论，并基于兰开斯特的特征值理论(Lancaster，1966)，是指产品或选择的效用不是由产品或选择本身提供，而是源自构成产品或选择的单一属性。相应地，效用和人们的偏好可以作为这些属性的线性函数。但是，根据随机效用理论，除了这种系统的组件，还有一个随机的组件(随机效用)，它呈现的是研究者没有观察到的其他偏好，以及判断中的错误和没有明确包括在研究者效用函数中的所有其他因子。这一随机的组件是随机效用理论很重要的一个方面，它将效用定义为一种潜在构念，只存在于人们的脑中，不能被研究者直接观察到(更多关于随机效用理论的信息，见Amaya-Amaya et al.，2008；Louviere et al.，2010；McFadden，1986)。因此，随机效用理论为分解调查法和回归法提供了很好的理论依据，例如那些在第5章提出的，包括L1误差项 ε_{ij}。大量实证研究支持了随机效用理论，并为差异集合研究和选择行为数据库提供了完美的框架(参见第6章第2节)。

以行为理论推动分析是选择性实验的一个优势。大多数因素调查研究也依赖于社会理论，但是选择性实验显示了使用这些理论来建构(回归)分析的模型是可取的。如第3章所述，这一策略提供了一种直接的方法来建立聪明而高效的实验设计。如果对于效用最大化和选择感兴趣，随机效用

理论可能为因素调查研究提供了一个很好的开端。

选择性实验的实验设计比因素调查的要求更多。这种差异的一个原因是使用了数据分析的非线性(logit 或概率)模型;这种模型意味着为了获得最大的效率,要对 *D*-efficiency 标准公式进行修改(详见第 3 章第 2 节)。[50] 此外,若干选项在一个选项集合中被绑定在一起,这导致了有效设计需要的两个额外标准,这超出了第 3 章第 2 节所描述的范围。首先,在一个选项集合中的选项属性水平之间应该有一个最小的重叠部分,让调查对象必须真正地将他们的选择建立在集中所有选项属性之上(而不仅仅是选项之间不同的属性)。其次,应该避免那些明显优于其他选项的占主导地位的选项。由于研究者可以提前预计调查对象的选择,所以这些选项只提供关于调查对象是否理解任务的信息,而没有提供关于他们真正偏好的信息。[51] 此外,最近的研究表明,如果主要的研究目标是对维度之间权衡的评估(比如支付意愿的估计和相关政策),我们可以通过测量设计性能来获得除了 *D*-efficiency 以外的效率(Scarpa & Rose, 2008)。我们建议读者,如果希望使用成对或多重比较作为应答任务,就请了解选择性实验设计中的这些特定文献(例如,R.F.Johnson et al., 2006; Louviere et al., 2000; Scarpa & Rose, 2008)。

选择性实验还产生了一个具有三个不同水平的分层数据结构。对单项选择(选或不选)的应答和它们的属性在 L1 中是不一样的,单个选择集合的标识表明哪些选项被绑定到一个选择上,在 L2 中是不同的,从而调查对象的特征在 L3 中有所不同。对于数据分析,可以使用常见的分析离散选择,但是我们应该解决多水平数据结构。这种数据结构源自

对每个调查对象的多重观察,是带有聚类标准误或多重分析的。如果在选择集合中只有两个选项,常常使用二元的 logit 或概率模型。然而,尤其在有标记的替代性方案的情况下,选择性实验可考虑有条件的 logit(conditional logit, CL)模型(Stata 的实用介绍请见 Long & Freese, 2006)。这些模型很容易估计和解释,但是它们依赖于无关备选方案的独立性假设(independence of irrelevant alternative, IIA)。[52]在这种假设没有得到满足的情况下,我们必须选择更复杂的模型,例如多个 logistic 回归或跨栏模型(hurdle models),它们可以对多层决策结构进行评估(例如嵌套 logit 模型)。这些模型可以在标准的计量学文献中找到。然而,选择性实验也有特别之处,例如使用替代性的特定常量。对使用选择任务感兴趣的读者(也可以在因素调查中实施)可能首先对选择性实验上的特定文献进行深入的测试(例如, Amaya-Amaya et al., 2008; Hensher et al., 2005; Louviere et al., 2000)。分析离散选择(例如, 有条件的 logit)的许多模型的一个限制是,调查对象变量对选择可能性的影响只能通过子集比较或跨水平交互作用来估计,不能通过主效应来估计。[53]社会科学中研究者对调查对象变量的影响有很强的兴趣,但这种方法可能会限制其在社会研究中的应用。

总之,如果研究者是对调查对象之间的差异感兴趣的话,选择性实验常用的数据分析技术的局限性是显而易见的。此外,设计选择性实验比设计因素调查更费时费事。对于大多数以了解态度、规范或定义为目的的研究而言,因素调查是更适合的方法。对于分析选择行为而言,选择性实验可能提供了一个更可靠的分析工具,因为它们得益于一个强

大的理论背景和对内外效度研究更多的方法论(Louviere et al., 2000;也可见第 6 章第 2 节)。然而,选择性实验仅限于选择任务。选择性实验的应答模式要求是让调查对象同时面对至少两种不同选项的表格式描述,然后被问及会选择哪一种。这一任务可能更倾向于社会期望偏见(social desirability bias, SDB),而不是因素调查研究中通常使用的段落描述。对调查对象而言,嵌入运行文本中的单一维度和实验操纵可能不那么突兀,但是在这一问题上的研究是匮乏的。此外,值得注意的是,在表格式呈现中减少社会期望偏见的技术方法是存在的(例如只使用在调查对象之间的差异,见第 6 章第 2 节)。

因此,决定是否使用一个典型的选择性实验或因素调查实验(或两者结合)的主要考虑是你希望使用哪种应答量表的类型。这个问题已经从方法论的角度在第 4 章第 2 节中讨论过。排序或选择任务迫使调查对象在不同选项中做出决定;因此,在这样的评估中,我们可能会得到比评分任务更大的变化。然而,当调查对象真的犹豫不决时,这些被迫的选择可能会导致对差异的过度估计或甚至完全武断的选择。应答删失的问题可以通过选择任务规避,但是我们只能用相对学术的语言来解释评估和推断结果(例如效用术语)。即使选择性实验的"不选择"选项提供了一些锚点,也无法对效用或偏好进行精确的测量。因此,对这两种应答量表都存在方法论的争论(评分和排序的比较,也可见 Alwin & Krosnick, 1985; Krosnick & Alwin, 1988)。

然而,因素调查一个可能是最重要的方面尚未被考虑。为了实现高水平的外部效度(就心理现实主义而言),我们可

能采用最切合研究的实质性主题的应答任务。在接下来的部分，我们将更详细地讨论外部效度的问题。在这一点上，可以说如果任务之间是强相关的话，“现实生活”行为和调查实验中揭示的行为之间的等价性就会增加。也就是说，当研究对象看起来“自然”时，行动者可以在同时呈现的选项中做出选择（例如，就像雇主做的许多雇用决策一样），我们可能会选择选择性实验方法。相反，在其他情况下（例如，研究按先后顺序进入不同工作机会的求职者的情况），我们可能会使用因素调查方法，因为它的实验设计简单，数据分析灵活。

小结

联合分析法和选择性实验是两类其他的调查实验。在这些实验中，调查对象面对的是一些属性层次不同的假设对象的描述。联合分析法通常作为这些分解调查实验的统称。然而，在更狭隘的意义上，市场营销研究中经常使用的联合分析法并不能满足许多学术研究的高标准。出于这些原因，因素调查和选择性实验应该被视为不同的方法，它们更适合于学术性的社会科学研究，而不仅仅是市场营销研究。

为了研究行为的意图，可考虑使用选择性实验而不是因素调查。与因素调查不同的是，选择性实验是专门为了测量不同对象或服务之间的选择而设计的，这些选择联合呈现在一个选择集合中。选择性实验的优点包括它们的由随机效应理论支撑的强理论框架、经济学理论偏好和能推动选择任务、回归技术和效度研究的效用最大化。然而，选择性实验的设计要求比因素调查多少更严苛一些，因为我们必须考虑

另一些标准来实现最有效的实验设计。此外，通常被用于选择性实验研究中的数据分析限制了测试调查对象间变量的可能性。除了这些标准外，选择什么方法取决于最适合当下研究对象的应答任务类型。研究者主要对不同选项之间的选择感兴趣，还是要求调查对象对不同选项进行单独评分更有益于研究？最重要的问题可能是，对于某些主题，在选择性实验（因素调查）中选择（评分）任务更“自然”，这有助于增加心理现实主义作为外部效度的部分。

第2节 因素调查结果的效度和推广度

对于因素调查几乎最严重的批评是研究者只能测量假设性的决策，而不是真实的决策。因素调查所测量的规范和态度有可能与调查对象的真实选择不同（如见 Collett & Childs, 2011; Eifler, 2007, 2010; Faia, 1980）。在接下来的部分，我们对关于因素调查的效度问题进行简短的总结，在奥斯普格和欣茨的论文中可以看到更详细的讨论（Auspurg & Hinz, 2013）。

在实验研究中，最重要的效度概念是内部效度，它是指结果变量的方差是否真的由实验操纵所导致的。如果实验场景产生了不在研究者的控制之下的系统性方差，那么内部效度就会受到威胁（Aronson et al., 1998; Brewer, 2000）。在因素调查研究中当不使用带有高解决度实验设计时，或者将虚拟情境随机分配给调查对象不成功时，或者当虚拟情境维度与调查对象的变量相混淆时，才可能出现这种情况。另一些威胁内部效度的情况包括水平的数量或排序的影响，或认知负荷和疲劳的影响等。因此，提升内部效度最重要的建议就是遵循第2章到第5章所提供的方法建议。

相比之下，构念效度（construct validity）指的是在一项研

究中能测到某一理论概念的程度。通过因素调查,我们只能尝试尽可能有效地测量假设性的决策、信念和意图。如果这些意图对真实行为的预测能力有限,就会影响到因素调查研究的相关性,而不是它的效度。构念效度的另一个威胁是社会期望偏差。该偏差可能主要是夸大社会期望行为的可能性(如失业人员接受工作机会),而可能不会对虚拟情境维度的因果影响产生偏差(如虚拟情境中描述的工作特征的影响)。对行为的普及性或可能性进行评估的最佳方式是使用非实验方法,因为实验不能确定真实世界中有多少人将会接触到这种处理方式(Mutz, 2011)。迄今为止,这个问题在许多效度研究中都被误解了(例如,见 Eifler, 2007, 2010)。那些已经充分讨论过实验刺激的因果影响的研究表明,与试图明显减少社会期望偏差的简单项目问题或甚至随机应答技术相比,因素调查较为有效(Armacost, Hosseini, Morris, & Rehbein, 1991; Auspurg et al., 2014; Mutz, 2011)(Armacost et al., 1991)。[54] SDB 的减少可能依赖于不引人注目的虚拟情境维度(Mutz, 2011)。此外,研究表明,基于他人的虚拟情境(即在虚拟情境中其他人行为的描述)导致承认敏感性行为的比率高于基于个人的虚拟情境(描述应答者自己的假设行为,见 Finch, 1987; Ganong/Coleman, 2006; Wason et al., 2002)。

最后的效度考虑的是外部效度,指的是对于不同样本的参与者、不同结果的测量,或其他实验、效应和它潜在的过程在其他情境中是否也可以被观察到(Brewer, 2000; Mutz, 2011)。结果不能被推广到其他情境的一个原因是缺乏内部效度或建构效度。因此要优先考虑内部效度和构念效度。

重要的是，当结果缺乏内在效度或构念效度时，了解其能否被推广到其他情境中是没有价值的。实验结果不能在其他情境中得到推广的原因只有一个——存在一些调节变量，它们在一种情境中起作用，而在其他情境中不起作用(Aronson et al., 1998; Campbell, 1957)。可能一些调节变量存在于维度中，但未被纳入虚拟情境。在现实世界中，这些调节变量可能会改善或阻碍因素调查研究中观察到的过程。当然，所有的实验实际上都在消除现实世界大部分的"噪音"。然而，这是确保高内在效度的关键。因此，这是一种优势，而不是限制(Brewer, 2000; Mutz, 2011)。

目前，人们对增强外部效度的方式知之甚少。就推广度而言，最重要的考虑可能是社会心理学家所谓的*心理现实主义*(psychological realism)，即在实验中出现的心理过程与现实生活中出现的相似程度(Aronson et al., 1998; Brewer, 2000)。推动这种心理现实主义是我们建议(在第6章第1节)选择代表当下研究问题的任务的主要原因。[55]然而，获得外部效度的理想手段是故意改变研究设计的各个方面，进行累积性研究。与其他所有调查实验类似，因素调查比实验室实验更有优势的地方在于，很容易测试不同样本人群或在不同地理位置上进行重复调查。这些变化是外部效度的良方；不同方法的互补性比任何单独的方法都更有助于提升外部效度(Brewer, 2000)。

为了检验*内部效度*，我们可以首先进行随机性检查(即确保虚拟情境维度和层块与调查对象变量无关)。其次，我们应该检测虚拟情境设计与现实样本之间是否具有统计学意义上的高效度。第三，我们建议检查可能出现的实验方法

问题，例如应答集合或认知负荷和疲劳的迹象。这种检测可以通过检查每个应答者虚拟情境评估的方差，检查评估的一致性来完成（例如解释性方差的数量或判断的信性），或分析调查对象倾向于通过忽视维度或其他形式的有问题的满意度来减少工作投入的线索（例如，见 Auspurg et al.，2009；Auspurg & Jäckle，2012；Sauer et al.，2011）。另外，不现实的短暂应答时间可能暗示着无效的应答。

构念效度的检测可以结合不同的应答量表或虚拟情境维度进行分签—投票设计，或者对理论概念中的其他数据进行交叉校验，如在虚拟情境维度中的直接项目问题（direct item question，如直接项目问题会问性别应该在多大程度上影响对收入公平的评估）。对同一概念的有效测量至少应该是相关的。测量社会期望偏差的一般策略（例如，见 Krumpal，2013）可以被因素调查研究采用，例如测试匿名程度是否会产生不同结果。

最后，研究外部效度的理想方式是比较不同的情境、方法和应答样本结果，从而确定结果是否在这些不同的条件下重复出现，更多了解其中有意义的调节变量。

比较实验方法和非实验方法的结果，我们应该牢记，所有这些研究都不可避免地测量了不同的现象。首先，由于它们较高的内部效度，实验比起其他方法通常有更强的统计检验力来确定因果关系。因此，我们必须看到实验数据分析比非实验数据标准误更低，影响更显著。其次，出于类似原因，无法解释的方差可能在不同的情境（或者不同调查对象样本）中有所变化。对于 logistic 或概率回归这样的非线性模型，这种不同的误差方差会产生很大的影响，因为它也会影

响回归系数的效应值(Mood, 2010)。因此,对于非线性模型,不应该直接比较斜率系数的绝对值。我们应该确定它是否等于一个合理的正常数(positive constant of propotionality,详见 Louviere et al., 2000; Mood, 2010)。最后,对于独立的回归模型(即线性或非线性),我们应该只比较斜率系数而不是截距来确认因果机制。截距测量了许多被忽略的数据来源间变化的变量的平均效应(Louviere, Hensher, & Swait, 2007),同时,实验和非实验数据肯定不同。

第3章和第4章中讨论了因素调查方法的内部效度和构念效度的研究(即方法问题的研究)。少数现存的对因素调查外部效度的研究关注的是因素调查研究中可观察到的行为意图间的一致性,以及通过观察研究、田野实验或人口调查测量"真实"行为的一致性(Eifler, 2007, 2010; Groß & Börensen, 2009; Nisic & Auspurg, 2009; Pager & Quillian, 2005)。在这些研究中,回归系数在不同情境和方法上的相对效应值通常是非常近似的,然而在截距中却存在显著的差异,相应地在行为选项的平均比例上也存在显著差距[例如,闯红灯或归还丢失的信件(会或不会)]。结果多少有些复杂,但他们主要显示了因素调查低估了社会不期望行为,并高估了社会期望行为。尽管如此,鉴于前几节的讨论,一些研究者认为因素调查"不能被视为有效的"(如 Eifler, 2010: 145)结论似乎为时过早。迄今为止,所有研究中(除了 Pager & Quillian, 2005),参与者各不相同。因此,结果的差异可能反映了样本构成的差异。到目前为止,不同环境下类似样本的研究结果喜忧参半。尼西克和奥斯普格(Nisic & Auspurg, 2009)关于出于工作原因移民意愿的研究得出的

结论是，大多数维度都对行为意图产生了类似的影响，这是由因素调查测量的；而真正移民是通过大规模的入户调查来调查和测量的。相比之下，佩吉尔和奎利恩(Pager & Quillian, 2005)关于招募决策的研究发现，基于同样的雇主样本，因素调查数据和田野实验数据之间存在巨大差异。不幸的是，这种差异在求职者的种族背景的因果关系中尤为明显，而这是研究的主要兴趣点。然而，正如作者所言，可能有许多原因导致行为和意图的差异。因此，需要更多的研究来解释这些差异。

尽管关于因素调查方法效度的研究仍处于早期阶段，但已经有大量关于其他调查方法如联合分析法和选择性实验等对效度的研究。这些研究涵盖了诸多不同的主题，如旅游模式选择、娱乐地区偏好、医疗服务，以及其他服务和产品的类别，但大多显示出结果的高一致性(一些综述性研究请见 Blamey & Bennett, 2001; Louviere et al., 2007; Louviere et al., 2010; Telser & Zweifel, 2007)。但是，由于应答任务和呈现方式的不同(见第 6 章第 1 节)，这些结果多大程度上可以被推广到因素调查尚不清楚。

小结

实验方法的主要优点是它们高内在效度。此外，构念效度是检验理论的关键。这两种类型的效度对于确保通过理论机制得到正确结论是必要的，这也是实验研究的主要兴趣；它们也是第三种效度(即外部效度)的前提条件。

为了提高内部效度，必须防止所有的方法论的人造结

果，这可以通过前几章提供的方法建议来实现。不同的虚拟情境维度或应答量表的分离有助于评估在不同操作化中的结果稳健性，这是构念效度的核心要素。此外，还可以根据因素调查以外的其他方法对结果进行交叉校验。在防止许多不合情理的虚拟情境下，应该能实现构念效度的第二个主要前提条件(即实验的现实主义)。确保外部效度和检验外部效度的理想方式是系统性地改变情境的特征和调查对象的样本。

在研究效度时，应该考虑与其他方法比较。实验方法具有更高的内在效度，因此带来更高的统计检验力。因此，应该将结果的比较限于回归系数，并排除基于标准误的参数，如显著性水平。对于所有的非线性回归模型，应该只检查斜率是否等于合理的正常数。

一些现有的因素调查外部效度的研究结论认为，因素调查缺乏有效性仅仅是因为描述结果的差异。鉴于健全的效度研究的不同概念和方法论标准，得出这一结论还为时过早。因果机制的结果(如斜率系数的比例)表明了因素调查方法是有效的，但结果不确定。正如其他相关实验方法(如选择性实验)的效度研究结果显示的，调查实验可以展现其高效度。



结　语

在许多社会科学的研究领域中，因素调查已经成为研究判断准则、社会规范、态度、定义和行为意图的一个成熟方法。基于实验设计的高内部效度，因素调查有信心为研究者提供在实验刺激和结果测量之间可观察到的联系是因果的结论。由此，这种方法非常适合于检验理论的研究。此外，因素调查还充分利用了调查研究的优势。因素调查相对轻松地覆盖不同地区的参与者的人口样本，或难以招募的实验研究对象。这种方法为分析中介变量或检验结果推广到不同背景和研究对象样本的普遍性提供了实质的可能性。换言之，因素调查具有高内部和外部效度的潜力。

在总结最重要的实践建议之前，本章的其余部分将提供关于在社会研究中因素调查与其他技术相结合的最新进展的信息。

第1节 | 与其他研究方法的结合

因素调查可以与质性访谈结合，有助于以更系统、更具说明性和情境化的形式规定谈话的主题（Finch，1987；Ganong & Coleman，2006）。在访问中，研究对象被要求为他们的虚拟情境评估提供理由（这可以通过使用封闭或开放的问题形式实现），或交流研究者没有在虚拟情境的必答题中考虑的其他问题，例如行为选择。这种在虚拟情境评估中更深层次的反思（可能包括其他人应该如何行为、可能的行为，或调查对象在虚拟情境场景中如何反应的问题）有助于丰富对理论概念的理解。此外，在定性和定量的访谈中，可以把虚拟情境故事分成不同的部分。这种策略被称为多部分因子虚拟情境设计（Ganong & Coleman，2006）。这种方法基本上在一份问卷中呈现了若干个相互关联的因素调查，研究者可以创建一个更复杂的虚拟情境故事，这个故事可以被分为不同部分。具体来说，在第一个虚拟情境中，介绍了主题（如年纪大的人需要帮助），并且研究对象在虚拟情境中（使用评分量表和/或开放式量表）回答问题。在随后的虚拟情境中，这个故事继续深化。调查者可以增加关于虚拟情境特征的信息（如，年长的人与主人公之间特定类型的关系），新的事件可能改变情境的框架（如对其中一个虚拟情境特征

进行长距离重置），或者描述一段时间后的场景（如，当虚拟情境主角变老或者大学毕业后）。这种方法使研究者可以观察到当描述的情况或特征的条件变化时，评估是如何变化的，这种改变在一个单独的虚拟情境中通常是很难描述的。此外，比起标准因素调查方法，这种方法可以在单一的问卷中测试大量不同的维度[维度的数量应该只有大约7（加减2）个]。我们也可以在不同的虚拟情境部分之后改变所要求的评估类型，这在标准的因素调查方法中是十分混乱的（详见 Ganong & Coleman，2006）。[56]然而，这种方法的一个关键缺点是，在每个部分中，研究对象只能对一个虚拟情境进行评分，否则，任务就会过于复杂。因此，我们必须招募更多的调查对象，这样，所有的虚拟情境都被评估或使用非常小的虚拟情境样本。在不同的部分间可能有顺序或疲劳效应。此外，有效的设计是复杂的，特别是因为要对每个调查对象的虚拟情境数量进行限制。[57]

因为在虚拟情境中描述的情境和对象比抽象的项目问题更具有说明性，所以它们可能被用于访谈儿童或认知能力相对较低的调查对象。据我们所知，目前为止，虚拟情境只被用于对儿童的访谈，因为它们在描述情境时运用类似故事的技术或在情境化的方式中提问题。虚拟情境被认为可以更好地吸引儿童的注意力，进而实现在研究过程中更高水平的控制，但是它们只在没有任何维度水平变化的情况下被使用（例如，见 Barter & Renold，2000）。然而，我们没有理由不采用实验变量。在这些儿童的研究中，我们必须使用一个比成人研究更低的复杂性水平（即更少的维度和虚拟情境）。关于哪种设计最适合哪个年龄段群体的问题，应该是未来实

证研究的目标。

一个有趣的方法发展是因素调查与实验室实验相结合(更深入的讨论见 Rauhut & Winter, 2010)。尽管因素调查方法可以研究复杂的规范或行为规则,但它仅限于假设性的评估。对因素调查的回答结果通常不会与参与者真实的结果相关。相反,参与者在实验中所做的决定与金钱的结果相关。然而,这种情况常常会受到只提供对规范或行为的单维度测量的缺点的影响(即在单维度刺激中的决定,是/否)。因此,在实验室实验中,研究者探索复杂模式的能力受到了限制,如条件或社会准则的强度。因素调查技术与实验室实验相结合则可以充分利用两者的优势。这样组合的一个例子是策略性方法(strategy method),调查对象首先被问及他们在不同假设的情况下如何反应,然后在随后的实验中设置其中一个情况(Fehr, Fischbacher, von Rosenbladt, Schupp, & Wagner, 2002; Rauhut & Winter, 2010)。因素调查方法的更多特性,比如更多可变的维度,可能对这种方法的进一步发展有积极作用。[58]

第 2 节 最重要的建议总结

遗憾的是，大部分现有的因素调查技术的应用或结合其他调查技术的应用都没有充分利用因素调查方法潜在的优势。许多使用者采用随机的虚拟情境样本，但忽略了实验设计中的统计研究文献，而且常常忽视用于多水平数据结构的高端技术分析工具（如多层回归）。[59]这种视而不见不仅会导致统计效率的损失，而且浪费资源（研究资金和调查时间），还会使基于因素调查数据的结论无效。同样地，许多研究者并不熟悉如何简化问卷编制或者节省管理多问卷版本的管理工作的技术（在第 4 章中有描述）。

基于这些原因，我们总结了七条我们认为对于获得尽可能有效、有统计效率和有启示的数据的最重要的建议。

1. 考虑调查对象的认知限制。给每个调查对象不超过 7（加减 2）个维度和 10 个虚拟情境，这样，不同年龄和教育水平的群体似乎都能完成要求繁重的评估任务，还能在没有认知负荷和疲劳效应的迹象下产生高水平的应答一致性。

2. 使用高设计分辨度（a high design resolution）的分数化的、*D*-efficient 虚拟情境样本。这些样本确保了我们可以识别出所有增益影响，这是所有因果解释的一个先决条件。此外，这样的样本允许统计效力的最大化（即相比效度较低

的设计，我们可以通过更少的评估或调查对象数量达到同样水平的统计效力）。此外，应该告知读者所使用的实验设计类型。

3. 确保虚拟情境和层块随机分配给调查对象。否则，实验设计的主要特征——即随机化——是实现不了的。所有虚拟情境应该被几个调查对象进行评分。随着调查对象样本的异质性增加，每个虚拟情境/层块的调查对象的数量要随之增加，以防止虚拟情境变量与调查对象变量的混淆问题。

4. 防止出现许多不合情理的虚拟情境。这样的虚拟情境削弱了实验的现实主义，触发了启发式的使用，从而导致了虚假性的人为结果。在建构实验设计时，应该考虑排除不合逻辑和不合情理的情况。如果这些情况在后续阶段才被清除，那么 *D*-efficient 样本的某些理想特性会被曲解，如维度的正交性。

5. 使用标准应答量表。诸如匹配数字技术的量级量表应该被放弃，应该支持更标准化的应答量表（至少在自我管理的调查中），从而实现高质量数据和低单位的无应答（low unit nonresponses）。对在同时呈现的选项间进行选择的真实生活情境中行为意图的应用，可以选择实验的应答任务和理论动机（随机效用理论）。

6. 考虑多水平数据结构，不使用两步数据分析技术。对每个单一调查对象进行多重评估的情况下，独立观察的假设是不成立的。尽管如此，聚类稳健标准误（调查对象个体的集群）或多水平分析可以正确估计标准误和显著性水平。两步骤方法通过对单一调查对象的观察进行单一回归，然后使

用一些个体特定的回归结果进一步分析，就存在不准确和不稳定的结果的劣势，具有较低的统计检验力。

7. 关注因果机制的内部效度。内部效度对于有价值的实验很重要。此外，对于理论验证性研究，构念效度应该优先于外部效度。因此，主要的建议是遵照因素调查设计的方法建议，并限制不同方法和样本之间的比较，以达到相似的斜率系数。对于非线性回归模型，应该只对比例常数进行测试。

上述建议不一定都能适用于所有的应用。例如，只有当虚拟情境全集的大小、调查对象的数量，以及调查可见的组合不允许对全集中的所有虚拟情境进行评估时，抽样技术才是必要的。类似地，当每个调查对象只使用一个虚拟情境时，不需要针对分层数据结果的特定数据分析技术。因此，对简单设计感兴趣的读者可能会跳过一些方法论的问题。然而，幸运的是，在节约研究资源的同时，例如调查对象样本的大小，目前的实验理论、统计的方法和电脑技术的状况允许建构具有高价值的内部效度和统计检验力的复杂设计。基于这些原因，我们相信在未来，因素调查将更频繁地被使用，并成为研究各种社会现象的一个有前景的工具。

注释

[1] 在第3章的第1节中详细解释了这八个维度的各个层次。

[2] 男性和女性角色差异有统计学意义($p<0.001$;通过随机截距回归对虚拟情境人物的收入和性别进行控制,并对所有调查对象进行计算)。

[3] 例如,2009年的德国成人样本中,月收入的"公平的性别收入差距"约为8%,换句话说,与女性员工相比,调查对象支持男性拥有平均8%更高的薪酬(Auspurg, Hinz, & Sauer, 2013)。

[4] 例如,收入公平性考察了调查对象多大程度上依赖于绩效、公平、需求的原则。为了分析这个问题,我们选择了与这些正义原则相关的维度(例如,工作经验指向每个员工在工作上的表现,而孩子的数量指向需求原则)。

[5] 这种随机变量可能源自一些(但不是全部)调查对象所采用的特殊性判断原则,或者是调查对象在判断中的出错或对虚拟情境案例增加了研究者没有预计的额外的假设。

[6] 对于选择实验的实验设计的其他案例,参见Louviere(1988),以及R.F. Johnson, Kanninen, and Özdemir(2006)的研究。对于使用相似维度的应用案例,见Jann(2005)。

[7] 设计分辨度背后的逻辑如下:如果一个设计是分辨度R,那么任何涉及$s<R$个因子的效应是不会与其他任何涉及最大数量$t=R-s-1$个因子的效应混淆的(Gunst & Mason, 1991:49)。例如,如果一个设计是分辨度Ⅴ,那么所有的双向交互作用(涉及$s=2$个因子)是不会与其他涉及最大数量$t=R-s-1=5-2-1=2$个因子的效应混淆的,这是双向交互作用(包括2个因子)或主效应(1个因子)。换言之,在分辨度Ⅴ的设计中,所有双向交互作用是可分离的,并且可以从所有主效应中分离出来。与此相反,分辨度Ⅳ设计为$t=R-s-1=4-2-1=1$。因此,双向交互作用不会与主效应相混淆,但是可能会与全部或是一些其他的双向交互作用相混淆。通常研究者对识别一些而不是全部双向交互作用感兴趣。在这些情况下,他们可以采用解决方案Ⅳ的设计,并将他们感兴趣的特定交互作用正交化,因此就不会被混淆。

[8] 卢维埃(Louviere, 1988:40)认为在社会科学中,三方交互作用很难解释超过2%—3%结果变量的解释辩论,而高阶交互作用只解释了差异的极小比例。

[9] 方差—协方差矩阵方程$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$,方差参数$\sigma^2$是一个未知的常数,可以被忽略(Kuhfeld, 2010)。D-efficiency是关于$\mathbf{X}'\mathbf{X}$矩阵几何均数的

函数。其他效率的衡量措施，如 *A*-或 *G*-efficiency 都是基于类似的数学概念，例如这个矩阵的算数平均数。总的来说，所有这些衡量措施都是紧密相关的。因此，只关注 *D*-efficiency 是很常见的。被用于比较不同设计的 *D*-efficiency 的价值比率，其优势在于不依赖用于维度的编码类型（例如正交或是效度编码）；因此，它们在不同背景下的可比性比其他测量措施更高（Kuhfeld，1997）。

[10] 因为不能确定我们能够找到最好的设计，建议倾向于选择 *D*-efficient 而不是 *D*-optimal 设计（Kuhfeld et al.，1994）。

[11] 这是通过交换水平组合以及采用不同的候选集合来实现的。

[12] http://support.sas.com/resources/papers/tnote/tnote_marketresearch.html.

[13] 正如海因穆勒、霍普金斯和山本（Hainmueller，Hopkins，and Yamamoto，2014）提出的，随机样本的另一个理由是，我们不需要提前说明估计模型，但是可以依赖非参数估计程序。然而，参数的识别仍然需要维度和交互条件的不相关，这只有通过非常大的虚拟情境随机分部才能实现。

[14] 例如，库菲尔德等人（Kuhfeld et al.，1994）为一个测试 1 000 个不同的随机样本的 3^{15} 设计，最好的随机样本只与最好的 *D*-efficient 设计的 57%一样有效。

[15] 这可以通过检查维度和交互条件的相关矩阵来执行。建议使用专业的软件包，例如 SAS macros，它显示了混淆的结构，进而识别出由几个维度引起的多重共线性。如果存在问题性的混淆，我们应该抛弃样本，测试新的样本。然而，这一行为与随机样本的统计理由相对（即它们完全是随机选择的）。当使用随机样本时，人们可能会使用分层抽样技术，这在概念上更类似于分部的抽样技术。另一种可能是完全随机抽样，每个调查对象都有自己的随机样本，这允许使用非常大的随机样本，因此也可以确保许多参数的可识别性（见 Hainmueller et al.，2014）。

[16] 例如，关于移动决策或消费者决策的研究，如是否购买二手车，已经使用了用概率或逻辑回归进行估计的二元回应（是否选择某选项）（例如，见 Buskens & Weesie，2000；Shlay，1985）。

[17] 当使用排序或选择任务作为回答量表时，存在另一些关于如何提高信息的建议。由于这些建议与抽样的统计理论无关，我们在第 6 章第 1 节讨论。

[18] SAS 宏%*Mktblock* 在保持水平平衡的同时，尽可能地在单个层块中保持正交性。从技术上讲，这是通过将层块编号（称为块）作为一个必正交化的实验因子来实现（Kuhfeld，2010）。

[19] 通过前面提到的 SAS 宏，这个目标的实现可以通过在本地宏（%*macro*

restrictions)中指定不满意("bad")的组合,然后使用%*Mktex* 宏的选项限制。%*Mktex* 宏将只在给定限制下搜寻最有效的设计。通过这种方法可能识别应该完全避免或在一定程度上避免(即未被充分代表)的组合。可以在我们的网站上找到具体的例子。

[20] 造成这种效率损失的原因是显而易见的。刻意排除的组合总是造成受影响维度(可能是进一步维度)间的相关性(可能是进一步维度),然后水平平衡也会受到影响。

[21] 请注意,只有适度排除的影响才能显示出来。当更多的组合被排除在外时,*D*-efficiency 就会呈指数下降。

[22] 另外,小的随机样本在 *D*-efficiency 上也会有较大的差异。例如,对于 50 个虚拟情境的 10 个随机样本的平均 *D*-efficiency 39.13 的标准差(第一行)是 4.11;对于排除在外的不合情理的情况的平均 *D*-efficiency 27.87(第二行),标准差更高(7.67)。SAS 算法的结果也显示一些随机变化,但是,它要小得多(分别是 0.04 和 0.12)。由于需要大量的计算能力,我们没有报告表 3.3 中所示的所有模拟值。然而,最重要的是,随机样本中的高变异量意味着低的标准化。因此,基于不同的随机样本的可比较性较低。

[23] 缺失观测的模版应该仔细检查。可能会出现不随机数据的丢失,这可能会导致参数估计的偏差和统计检验力的降低。

[24] 此外,比例公式(方程 3.5)只适用于大数量的案例($n>30$),因为它是基于渐近正态性的二项分布。

[25] 换言之,组内相关 ρ 表明在结果中变量的多少是由不同调查对象估计虚拟情境引起的。对于方程 3.7,这个系数被假定是大于零。一般来说,调查对象样本变得更加异质化,组内相关就会增加。例如,在调查普通人群收入公平的例子中(详见第 2 章),组内相关为 0.12 左右(参见第 5 章第 2 节)。对于收入公平的类似因素调查,是由德国大学生进行的(呈现更加同质化的调查对象样本),组内相关只有大概 0.06。

[26] 对于回归系数 β 为零的零假设,可以使用以下公式计算出 SAS 中的双边测试的检验力,这里 se($\hat{\beta}$)表示估计值 $\hat{\beta}$ 的标准误,z_x 是标准正态分布的分位数,与 $p=x$ 的概率一致,α 是显著性水平,而 b 是检验力(Snijder, 2005:3):$\frac{\beta}{se(\beta)} \approx z_{1-\frac{\alpha}{2}} + z_b$。这个公式可以应用于检验力 b 至少为 0.03 的所有值。举个例子,使用显著性水平 $\alpha=0.05$ 并获得至少 0.08 的检验力,关键值为 $z_{1-\frac{\alpha}{2}}=1.96$ and $z_b=0.84$。因此,真实的参数值应该至少是标准误的 2.8 倍。

[27] 根据我们的经验,带有调查对象(不是虚拟情境)数量作为样本容量的

SRSs 的简单公式为我们期望的检验力提供了一个粗略的估计。

[28] 例如，你可以使用免费软件电脑程序 PinT，它是由汤姆·斯尼德斯、罗尔·博斯克和亨克·古尔德蒙德(Tom Snijder, Roel Bosker and Henk Guldemond)写的(更多信息见 http://www.stats.ox.ac.uk/~snijders/multilevel.htm)。

[29] 或者我们可以再次使用相关的程序 PinT。

[30] 最近的一项研究发现，在"就业歧视定义"上三个不同的因素调查模块有非常相似的结论，它们都使用了人口抽样样本，而且"众包招募"(crowdsource recruitment)调查成本比较低，但结果也较难推广(见 Weinberg, Freese, & McElhattan, 2014)。

[31] 即数据分析中，所有的判断在使用前通过参照虚拟情境判断被标明刻度(划分)。所有调查对象都需要额外的虚拟情境。

[32] 当使用特定调查对象分析时，例如每个调查对象个体单独进行的回归，我们不一定需要量级评估标刻度。然而，这些分析技术有许多其他缺点，因此我们不推荐(详见第 5 章)。

[33] 例如，由于在虚拟情境中一般都是更高(更低)的收入，锚定效应可能存在于调查对象评价公平收入的更高(更低)金额。当要求调查对象指出公平收入的数量时，我们可以避免在虚拟情境中标注任何数量，这样就可以避免调查对象使用虚拟情境中列举的数量作为判断的标准。

[34] 这些微小的操作可以通过观看克里森等人(Krysan et al., 2009)的视频虚拟情境来体验，在下面的网站中提供：http://www.psc.isr.umich.edu/tmp/das/。

[35] 然而，这种关联也可能由于文本虚拟情境中人物的特定命名引起(例如，见 Mutz, 2011)。为了避免这种关联，建议使用普通名字。

[36] 电话访问可能为调查对象提供通过混合访问方式来回顾虚拟情境，其中因素调查作为一种问卷先发送给调查对象。然后，访问员可以在电话访问中记录调查对象对虚拟情境的评估(假设调查对象在访问前完成了模块)，或调查对象可以用电子邮件将问卷寄回调查机构，调查机构会记录评估结果。后一种情况下，我们必须确保在这两种调查模式中对调查对象进行明确和独特的身份标记。

[37] 当我们只对最终调查对象的数量进行粗略的预测时，我们必须非常保守，并确保有足够数量的问卷版本提供给调查对象。在这样的情况下，许多问卷版本在建立数据库的最后没有被用上，进而没有将所有问卷版本完全随机分配可能会产生更平衡的应答数据，并产生更大的统计效率。因此要使用一个层块的结构，所有呈现一个虚拟情境样本的 n_i 虚拟情境被一起放入层块。然后，不同的层块随之附加到数据矩阵上

成为初始设置数据。因此，当所有其他层块已经被呈现至少 $k-1$ 次，我们可以确保每个单一的层块只呈现第 k 次。换言之，就单个虚拟情境层块来说，我们可以确保结果数据最大化平衡。但是在使用这种技术时，我们必须确保所有层块被随机分配给调查对象。

[38] 然而，请注意，在排除了不合逻辑的案例之后，在虚拟情境维度间将会产生相关性。在这种情况下，只有多元分析才能确保单维度影响被正确评估。

[39] 此外，这种方法没有考虑到第二阶段的结果变量(即斜率)表示估计，这可能导致错误的结论。

[40] 本质上，不同层块代表第三个水平(L3)，但包括对不同层块的虚拟变量，它实际上与带有 L3(层块)变量的三层次模型的估计是一样的。

[41] 这个限制是由于调查对象变量被完全混淆于特定调查对象的截距("固定效应")，因此无法识别。

[42] 另一个包含虚拟变量的层块并没有显示出任何显著的层块效应，对于虚拟情境变量结果也没有实质性的变化。

[43] $\rho = \mathrm{Var}(u_j)/[\mathrm{Var}(u_j)+\mathrm{Var}(\varepsilon_{ij})] = \mathrm{Std\ Dev}^2(u_j)/[\mathrm{Std\ Dev}^2(u_j)+\mathrm{Std\ Dev}^2(\varepsilon_{ij})]$.

[44] 固定效应回归模型产生的数据头(*xtreg*, *fe*)显示了与预测变量相关性 u_j。在我们的例子中，这个相关性使 $r=-0.011\ 9$。这个值接近 0，表明成功随机化。

[45] 但是，请注意，分类变量的影响也取决于类别的选择。我们可能使用更复杂的方法测量相对权重，也可以对相关回归系数进行正确估计(J.W. Johnson, 2000; Soofi, Retzer, & Yasai-Ardekani, 2000)。

[46] 研究者常常计算观察结果的边际效应，即关于自变量边际(一单元)变化的条件均值的估计值的影响(Cameron & Truvedi, 2009:527)。但是，这些值是次要的，我们通常希望能预测未观察到的和删失的变量。

[47] 在一个研究中允许使用更多属性的另一个方法是具有*分层*的多水平联合分析法，它使用两个及以上带有不同属性的资料卡的子样本。在所有的子样本中至少使用一个维度。这个"桥"保证了基于不同子样本的结果的比较。在桥接维度(权衡)的相对重要性方面，包含在不同子样本中的属性找到了一个通用的度量标准。这种方法可能对因素调查研究具有吸引力，但对于这种方法的有效解释依赖于几个假设(例如，权衡不是由进一步的维度来调节的)。

[48] 这些结果可以用来告知参与公共事务管理和社会政策的人。此外，这种假设的评估所提供的信息可用于法庭案件，以确定犯罪造成的生态损害，从而确定惩罚的程度(例如，在 Exxon *Valdez* 石油泄漏事件中就

是必要的)。

[49] 在设计选择性实验时,另一个有趣的考虑是是否使用标签的或去标签(一般的)的选择方案。例如,在研究选择旅游模式的决策时,人们可能会将不同的选项标记为火车、巴士或小汽车,或选项 1、选项 2 和选项 3。当一个人主要对单一属性间的影响和权衡(例如旅行时间和旅行成本)感兴趣,对不同旅行模式本身不那么感兴趣时,我们建议选择后者作为变量(即一般的选择);否则,这些标签可能会很强烈地甚至完全支配调查对象的选择(Amaya-Amaya et al., 2008:22)。

[50] 虽然效用被认为是属性的线性函数,但对于选择在数据分析中被估计的单个选项的概率而言,这不再是正确的(详见 Amaya-Amaya et al., 2008; Hensher et al., 2005; Louviere et al., 2000),这需要对 *D*-efficiency 的标准公式进行调整(见第 3 章第 2 节)。

[51] 一些作者建议应该使用*效用平衡*——即我们不应该只避免占主导地位的选项,还要试图将这些选项组合进单个选择集合中,这个选择集合要尽可能与它们的效用水平相似(Huber & Zwerina, 1996)。但是,当调查对象对选项漠不关心时,这可能会引发完全武断的选择(Louviere et al., 2008)。

[52] 根据该假设,选择 A 选项或 B 选项的可能性仅仅取决于这两个选项,因此,它们不会随着加入的 C 选项而改变。

[53] 这个限制是因为主要效果与有条件的 logit 模型中代表的固定效果的单项选择集合的标识完全混淆。

[54] 正如穆茨(Mutz, 2011)所指出,因素调查比起实验室研究,不太倾向于社会期望偏差,因为参与者对实验的参与意识较低,相应地,当进入实验室时,他们经验处理的意识也较低。

[55] 然而,这个建议是基于理论推理而不是实证研究。如前所述,考虑到目前的实证情况,还不清楚评分或者任务选择是否可以达到更高水平的数据质量(包括内部效度、构念效度和外部效度)。

[56] 为了提供更合理的序列或为了"让调查对象更多地参与到故事中",它可能有助于决定随后的回答部分,这是在之前阶段就提供的(Finch, 1987)。这一目标可以通过访谈软件工具来实现(移动决策的一个应用,见 Li et al., 2007)。然而,这种方法可能会丧失所有调查对象的可比性。

[57] 到目前为止,在用法说明上,有效设计的问题被忽视了,错误的随机样本作为黄金标准被推荐给因素调查实验。

[58] 此外,关于不同评价风格和参考点,锚定虚拟情境有助于标准化不同调查对象的发现。正如第 2 章指出的,与因素调查相比,设计锚定虚拟情

境是为了追求不同的研究目的。尽管如此，两种方法有几个共同的特点。考虑到这两种方法的优势，或者在单一的调查研究中结合这两种方法，将有助于丰富因素调查的内部效度和锚定虚拟情境。

[59] 例如，经常宣称虚拟情境维度的随机选择确保实验因子相互正交性(例如，见 Ganong & Coleman，2006：462)；然而，这一结果并没有得到保证，而且可能性比使用因子实验设计时更低。

参考文献

Abraham, M., Auspurg, K., Bähr, S., Frodermann, C., Gundert, S., & Hinz, T. (2013). Unemployment and the willingness to accept job offers: First results from a factorial survey approach. *Journal of Labour Market Research, 46*(4), 283–305.

Abraham, M., Auspurg, K., & Hinz, T. (2010). Migration decisions within dual-earner partnerships: A test of bargaining theory. *Journal of Marriage and Family, 72*(4), 876–892.

Adamowicz, W., Boxall, P., Williams, M., & Louviere, J. J. (1998). Stated preference approaches for measuring passive use values: Choice experiments and contingent valuation. *American Journal of Agricultural Economics, 80*(1), 64–75.

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly, 42*(1), 93–104.

Allison, P. D. (2009). *Fixed effects regression models*. Thousand Oaks, CA: Sage.

Alriksson, S., & Öberg, T. (2008). Conjoint analysis for environmental evaluation. *Environmental Science and Pollution Research, 15*(3), 244–257.

Alves, W. M., & Rossi, P. H. (1978). Who should get what? Fairness judgments of the distribution of earnings. *American Journal of Sociology, 84*(3), 541–564.

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly, 48*(4), 535–552.

Amaya-Amaya, M., Gerard, K., & Ryan, M. (2008). Discrete choice experiments in a nutshell. In M. Ryan, K. Gerard, & M. Amaya-Amaya (Eds.), *Using discrete choice experiments to value health and health care* (pp. 13–46). Dordrecht, the Netherlands: Springer.

Armacost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information. *Decision Sciences, 22*(5), 1073–1090.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 99–142). New York: McGraw-Hill.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 128–138.

Auspurg, K., & Hinz, T. (2013). *Validity and generalizability of factorial survey research: A comment* (Working paper of the DFG-Project "The Factorial Survey as a Method for Measuring Attitutes in Population Surveys" 11). Konstanz, Germany: University of Konstanz.

Auspurg, K., Hinz, T., & Liebig, S. (2009, August). *Complexity, learning effects and plausibility of vignettes in factorial surveys*. Paper presented at the 104th Annual Meeting of the American Sociological Association (ASA), San Francisco, CA.

Auspurg, K., Hinz, T., & Sauer, C. (2013, August). *Status construction or statistical discrimination? New insights on fair earnings from a factorial survey study*. Paper presented at the 108th Annual Meeting of the American Sociological Association (ASA), New York, NY.

Auspurg, K., Hinz, T., Sauer, C., & Liebig, S. (2014). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: Lessons from recent research*. (pp. 137–149). New York: Routledge.

Auspurg, K., & Jäckle, A. (2012). *First equals most important? Order effects in vignette-based measurement* (ISER Working Paper 2012-1). Essex, UK: Institute for Social & Economic Research (ISER).

Barter, C., & Renold, E. (2000). "I wanna tell you a story": Exploring the application of vignettes in qualitative research with children and young people. *International Journal of Social Research Methodology, 3*(4), 307–323.

Beck, M., & Opp, K.-D. (2001). Der faktorielle Survey und die Messung von Normen [The factorial survey and the measuring of norms]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 53*(2), 283–306.

Bennett, J., & Adamowicz, V. (2001). Some fundamentals of environmental choice modelling. In J. Bennett & R. Blamey (Eds.), *The choice modelling approach to environmental valuation* (pp. 37–72). Cheltenham, UK: Edward Elgar.

Berk, R. A., & Rossi, P. H. (1977). *Prison reform and state elites*. Cambridge, MA: Ballinger.

Blamey, R. K., & Bennett, J. (2001). Yea-saying and validation of a choice model of green product choice. In J. Bennett & R. K. Blamey (Eds.), *The choice modelling approach to environmental valuation* (pp. 178–201). Cheltenham, UK: Edward Elgar.

Brewer, M. B. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16). Cambridge, UK: Cambridge University Press.

Burke, W. J. (2009). Fitting and interpreting Cragg's tobit alternative using Stata. *Stata Journal, 9*(4), 584–592.

Buskens, V., & Weesie, J. (2000). An experiment on the effects of embeddedness in trust situations. *Rationality and Society, 12*(2), 227–253.

Byers, B., & Zeller, R. A. (1998). Measuring subgroup variation in social judgment research: A factorial survey approach. *Social Science Research, 27*(1), 73–84.

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using Stata*. College Station, TX: Stata Press.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312.

Carson, R. T., Louviere, J. J., & Wasi, N. (2009). A cautionary note on designing discrete choice experiments: A comment on Lusk and Norwood's "effect of experiment design on choice-based conjoint valuation estimates". *American Journal of Agricultural Economics, 91*(4), 1056–1063.

Champ, P. A., & Welsh, M. P. (2006). Survey methodologies for stated choice studies. In B. J. Kanninen (Ed.), *Valuing environmental amenities using stated choice studies: A common sense approach to theory and practice* (pp. 21–42). Dordrecht, the Netherlands: Springer.

Charlton, J. (2002). Review: Factorial survey methods: A valuable but under-utilised research method in nursing research? *Nursing Times Research, 7*(1), 44–45.

Chrzan, K., & Orme, B. (2000). *An overview and comparison of design strategies for choice-based conjoint analysis*. Sequim, WA: Sawtooth.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Collett, J. L., & Childs, E. (2011). Minding the gap: Meaning, affect, and the potential shortcomings of vignettes. *Social Science Research, 40*(2), 513–522.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica, 39*(5), 829–844.

de Wolf, I., & van der Velden, R. (2001). Selection processes for three types of academic jobs: An experiment among Dutch employers of social sciences graduates. *European Sociological Review, 17*(3), 317–330.

Diefenbach, H., & Opp, K.-D. (2007). When and why do people think there should be a divorce? An application of the factorial survey. *Rationality and Society, 19*(4), 485–517.

Dülmer, H. (2007). Experimental plans in factorial surveys: Random or quota design? *Sociological Methods and Research, 35*(3), 382–409.

Eifler, S. (2007). Evaluating the validity of self-reported deviant behavior using vignette analyses. *Quality and Quantity, 41*(2), 303–318.

Eifler, S. (2010). Validity of a factorial survey approach to the analysis of criminal behavior. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 139–146.

Emerson, M. O., Yancey, G., & Chai, K. J. (2001). Does race matter in residential segregation? Exploring the preferences of White Americans. *American Sociological Review, 66*(6), 922–935.

Faia, M. (1980). The vagaries of the vignette world: A comment on Alves and Rossi. *American Journal of Sociology, 85*(1), 951–954.

Fehr, E., Fischbacher, U., Von Rosenbladt, B., Schupp, J., & Wagner, G. (2002). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Journal of Applied Social Science Studies [Schmollers Jahrbuch], 122*(4), 519–542.

Ferrini, S., & Scarpa, R. (2007). Designs with apriori information for nonmarket valuation with choice-experiments: A Monte Carlo study. *Journal of Environmental Economics and Management, 53*(3), 342–262.

Finch, J. (1987). The vignette technique in survey research. *Sociology, 21*(1), 105–114.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Los Angeles, CA: Sage.

Franzen, A., & Pointner, S. (2012). Anonymity in the dictator game revisited. *Journal of Economic Behavior and Organization, 81*(1), 74–81.

Ganong, L. H., & Coleman, M. (2006). Multiple segment factorial vignette designs. *Journal of Marriage and Family, 68*(2), 455–468.

Garret, K. (1982). Child abuse: Problems of definition. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments. The factorial survey approach* (pp. 177–204). Beverly Hills, CA: Sage.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing, 54*(4), 3–19.

Greene, W. H. (2003). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Groß, J., & Börensen, C. (2009). Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung [How valid are measurements of behavior by means of vignettes? A methodological comparison of factorial survey and observational data]. In P. Kriwy, C. Gross, & M. Jungbauer-Gans (Eds.), *Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen* (pp. 149–178). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. Hoboken, NJ: John Wiley.

Gunst, R. F., & Mason, R. L. (1991). *How to construct fractional factorial experiments*. Milwaukee, WI: ASQC Quality Press.

Gustafsson, A., Herrmann, A., & Huber, F. (2007). Conjoint analysis as an instrument of market research practice. In A. Gustafsson, A. Herrmann, & F. Huber (Eds.), *Conjoint measurement: Methods and applications* (4th ed., pp. 3–30). Berlin, Germany: Springer.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251–1271.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014): Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis,* 22 (1), 1–30.

Hechter, M., Ranger-Moore, J., Jasso, G., & Horne, C. (1999). Do values matter? An analysis of advance directives for medical treatment. *European Sociological Review, 15*(4), 405–430.

Hembroff, L. A. (1987). The seriousness of acts and social contexts: A test of Black's theory of the behavior of law. *American Journal of Sociology, 93*(2), 322–347.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.

Hensher, D. A., Rose, J. M., & Greene, W. H. (2005). *Applied choice analysis*. Cambridge, UK: Cambridge University Press.

Hermkens, P. L. J., & Boerman, F. A. (1989). Consensus with respect to the fairness of incomes: Differences between social groups. *Social Justice Research, 3*(3), 201–215.

Hole, A. R. (2007). A comparison of approaches to estimating confidence intervals for willingness to pay measures. *Health Economics, 16*(8), 827–840.

Horne, C. (2003). The internal enforcement of norms. *European Sociological Review, 19*(4), 335–343.

Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The analysis of factorial surveys. *Sociological Methods and Research, 19*(4), 493–510.

Huber, J., & Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research, 33*(3), 307–317.

Jann, B. (2005). *Erwerbsarbeit, Einkommen und Geschlecht. Studien zum Schweizer Arbeitsmarkt* [*Industriousness, income and gender: Studies regarding the Swiss labor market*]. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Jasso, G. (1988). Whom shall we welcome? Elite judgments of the criteria for the selection of immigrants. *American Sociological Review, 53*(6), 919–932.

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods and Research, 34*(3), 334–423.

Jasso, G. (2012). Safeguarding justice research. *Sociological Methods and Research, 41*(1), 217–239.

Jasso, G., & Opp, K.-D. (1997). Probing the character of norms: A factorial survey analysis of the norms of political action. *American Sociological Review, 62*(6), 947–964.

Jasso, G., & Rossi, P. H. (1977). Distributive justice and earned income. *American Sociological Review, 42*(4), 639–651.

Jasso, G., & Webster, M., Jr. (1997). Double standards in just earnings for male and female workers. *Social Psychology Quarterly, 60*(1), 66–78.

Jasso, G., & Webster, M., Jr. (1999). Assessing the gender gap in just earnings and its underlying mechanisms. *Social Psychology Quarterly, 62*(4), 367–380.

John, C. S., & Bates, N. A. (1990). Racial composition and neighborhood evaluation. *Social Science Research, 19*(1), 47–61.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35*(1), 1–19.

Johnson, R. F., Kanninen, B., Bingham, M., & Özdemir, S. (2006). Experimental design for stated choice studies. In B. Kanninen (Ed.), *Valuing environmental amenities using stated choice studies: A common sense approach to theory and practice* (pp. 159–202). Dordrecht, the Netherlands: Springer.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191–207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis, 15*(1), 46–66.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201–219.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52*(4), 526–538.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity, 47*(4), 2025–2047.

Krysan, M., Couper, M. P., Farley, R., & Forman, T. A. (2009). Does race matter in neighborhood preferences? Results from a video experiment. *American Journal of Sociology, 115*(2), 527–559.

Kuhfeld, W. F. (1997). *Efficient experimental designs using computerized searches*. Sequim, WA: SAS Institute.

Kuhfeld, W. F. (2010). *Marketing research methods in SAS: Experimental design, choice, conjoint and graphical techniques*. Cary, NC: SAS Institute.

Kuhfeld, W. F., Tobias, R. D., & Garrat, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research, 31*(4), 545–557.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*(2), 132–157.

Li, J.-C. A., Chang, E. C., & Jasso, G. (2007, August). *Computerized multivariate factorial survey*. Paper presented at the 102nd Annual Meeting of the American Sociological Association, New York, NY.

Liebig, S., & Mau, S. (2005). Wann ist ein Steuersystem gerecht? Einstellungen zu allgemeinen Prinzipien der Besteuerung und zur Gerechtigkeit der eigenen Steuerlast [When is a tax system just? Attitudes towards general principles of taxation and the justice of tax burdens]. *Zeitschrift für Soziologie, 34*(6), 468–491.

Lodge, M., & Tursky, B. (1981). On the magnitude scaling of political opinion in survey research. *American Journal of Political Science, 25*(2), 376–419.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). College Station, TX: Stata Press.

Louviere, J. J. (1988). *Analyzing decision making: Metric conjoint analysis*. Newbury Park, CA: Sage.

Louviere, J. J. (2006). What you don't know might hurt you: Some unresolved issues in the design and analysis of discrete choice experiments. *Environmental and Resource Economics, 34*(1), 173–188.

Louviere, J. J., Flynn, T. N., & Carson, R. T. (2010). Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling, 3*(3), 57–72.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: Analysis and application*. Cambridge, UK: Cambridge University Press.

Louviere, J. J., Hensher, D. A., & Swait, J. (2007). Conjoint preference elicitation methods in the broader context of random utility theory preference elicitation methods. In A. Gustafsson, A. Herrmann, & F. Huber (Eds.), *Conjoint measurement: Methods and applications* (4th ed., pp. 167–198). Berlin, Germany: Springer.

Louviere, J. J., Islam, T., Wasi, N., Street, D., & Burgess, L. (2008). Designing discrete choice experiments: Do optimal designs come at a price? *Journal of Consumer Research, 35*(2), 360–375.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1–27.

Ludwick, R., Wright, M. E., Zeller, R. A., Dowding, D. W., Lauder, W., & Winchell, J. (2004). An improved methodology for advancing nursing research: Factorial surveys. *Advances in Nursing Science, 27*(3), 224–238.

Ludwick, R., & Zeller, R. A. (2001). The factorial survey: An experimental method to replicate real world problems. *Nursing Research, 50*(2), 129–133.

Markovsky, B., & Eriksson, K. (2012). Comparing direct and indirect measures of just rewards: What have we learned? *Sociological Methods and Research, 41*(1), 240–245.

McDermott, R. (2002). Experimental methodology in political science. *Political Analysis, 10*(4), 325–342.

McFadden, D. (1976). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York, NY: Academic Press.

McFadden, D. (1986). The choice theory approach to market research. *Marketing Science, 5*(4), 275–297.

Meudell, B. M. (1982). Household and social standing: Dynamic and static dimensions. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach* (pp. 69–94). Beverly Hills, CA: Sage.

Meyerhoff, J., & Liebe, U. (2009). Status quo effect in choice experiments: Empirical evidence on attitudes and choice task complexity. *Land Economics, 85*(3), 515–528.

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 101*(2), 343–352.

Miller, J. L., Rossi, P. H., & Simpson, J. E. (1986). Perceptions of justice: Race and gender differences in judgments of appropriate prison sentences. *Law and Society Review, 20*(3), 313–334.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67–82.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton, NJ: Princeton University Press.

Nisic, N., & Auspurg, K. (2009). Faktorieller Survey und klassische Bevölkerungsumfrage im Vergleich: Validität, Grenzen und Möglichkeiten beider Ansätze [Comparison of factorial surveys and classic population surveys: Validity, limits and potential of both methods]. In P. Kriwy & C. Gross (Eds.), *Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen* (pp. 211–246). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Nock, S. L. (1982). Family social standing: Consensus on characteristics. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach* (pp. 95–118). Beverly Hills, CA: Sage.

O'Toole, R., Webster, S. W., O'Toole, A. W., & Lucal, B. (1999). Teachers' recognition and reporting of child abuse: A factorial survey. *Child Abuse and Neglect, 23*(11), 1083–1101.

Pager, D., & Quillian, L. (2005). Walking the talk? What employers say versus what they do. *American Sociological Review, 70*(3), 355–380.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata* (2nd ed.). College Station, TX: Stata Press.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Rauhut, H., & Winter, F. (2010). A sociological perspective on measuring social norms by means of strategy method experiments. *Social Science Research, 39*(6), 1181–1194.

Rogers, W. H. (1994). Regression standard errors in clustered samples. *Stata Technical Bulletin, 3*(13), 19–23.

Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research: Papers in honor of Paul F. Lazarsfeld* (pp. 176–186). New York, NY: Free Press.

Rossi, P. H., & Alves, W. M. (1980). Rejoinder to Faia. *American Journal of Sociology, 85*(4), 954–955.

Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach* (pp. 15–67). Beverly Hills, CA: Sage.

Rossi, P. H., Sampson, W. A., Bose, C. E., Jasso, G., & Passel, J. (1974). Measuring household social standing. *Social Science Research, 3*(3), 169–190.

Ryan, M., Gerard, K., & Amaya-Amaya, M. (2008). *Discrete choice experiments to value health and health care*. Dordrecht, the Netherlands: Springer.

Ryan, M., McIntosh E., Dean, T., & Old. P. (2000): Trade-offs between location and waiting times in the provision of health care: the case of elective surgery on the Isle of Wight. *Journal of Public Health,* 22(2), 202–210.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty, 1*(1), 7–59.

Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2011). The application of factorial surveys in general populations samples: The effects of respondent age and education on response times and response consistency. *Survey Research Methods, 5*(3), 89–102.

Sauer, C., Auspurg, K., Hinz, T., Liebig, S., & Schupp, J. (2014). *Method effects in factorial surveys: An analysis of respondents' comments, interviewers' assessments, and response behavior* (SOEP Papers on Multidisciplinary Panel Data Research 629). Berlin: German Institute for Economic Research (DIW).

Scarpa, R., & Rose, J. M. (2008). Design efficiency for non-market valuation with choice modelling: How to measure it, what to report and why. *Australian Journal of Agricultural and Resource Economics, 52*(3), 253–282.

Schaeffer, N. C., & Bradburn, N. M. (1989). Respondent behavior in magnitude estimation. *Journal of the American Statistical Association, 84*(406), 402–413.

Schrenker, M. (2009). Warum fast alle das deutsche Rentensystem ungerecht finden, aber trotzdem nichts daran ändern möchten [Why almost everybody considers the German pensions system to be unjust, but nobody wants to change it: The perception of just pensions and the acceptance of reforms]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 61*(2), 1–24.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording and context.* New York, NY: Academic Press.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*(2), 277–287.

Schwarz, N., & Knäuper, B. (2000). Cognition, aging and self-reports. In D. C. Park & N. Schwarz (Eds.), *Cognitive aging: A primer* (pp. 233–252). Philadelphia, PA: Psychology Press.

Shepelak, N. J., & Alwin, D. F. (1986). Beliefs about inequality and perceptions of distributive justice. *American Sociological Review, 51*(1), 30–46.

Shlay, A. B. (1985). Castles in the sky: Measuring housing and neighborhood ideology. *Environment and Behavior, 17*(5), 593–626.

Shlay, A. B. (1986). Taking apart the American dream: The influence of income and family composition on residential evaluations. *Urban Studies, 23*(4), 253–270.

Sniderman, P. M., & Grob, D. B. (1996). Innovations in experimental design in attitude surveys. *Annual Review of Sociology, 22,* 377–399.

Snijders, T. A. B. (2001). Sampling. In A. Leyland & H. Goldstein (Eds.), *Contribution to multilevel modelling of health statistics* (pp. 159–174). Chichester, UK: Wiley.

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1570–1573). Chichester, UK: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*(3), 237–259.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.

Soofi, E. S., Retzer, J. J., & Yasai-Ardekani, M. (2000). A framework for measuring the importance of variables with applications to management research and decision models. *Decision Sciences, 31*(3), 595–625.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153–181.

Struck, O., Krause, A., & Pfeifer, C. (2008). Entlassungen: Gerechtigkeitsempfinden und Folgewirkungen [Dismissals: Feelings of justice and subsequent effects. Theoretical concepts and empirical results]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 60*(1), 106–126.

Telser, H., & Zweifel, P. (2007). Validity of discrete-choice experiments evidence for health risk reduction. *Applied Economics, 39*(1), 69–78.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*(1), 24–36.

Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'Andrade, R. (1989). Carryover effects in attitude surveys. *Public Opinion Quarterly, 53*(4), 495–524.

Vonesh, E. F., & Chinchilli, V. G. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. London, UK: Chapman & Hall.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research, 38*(3), 505–520.

Wallander, L. (2012). Measuring social workers' judgements: Why and how to use the factorial survey approach in the study of professional judgements. *Journal of Social Work, 12*(4), 364–384.

Wason, K. D., Polonsky, M. J., & Hyman, M. R. (2002). Designing vignette studies in marketing. *Australasian Marketing Journal, 10*(3), 41–58.

Weinberg, J., Freese, J., & McElhattan, D. (2014). Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. *Sociological Science, 1*, 292–310.

Will, J. A. (1993). The dimensions of poverty: Public perceptions of the deserving poor. *Social Science Research, 22*(3), 312–332.

Wirtz, J. (1996). Controlling halo in attribute-specific customer satisfaction measures: Towards a conceptual framework. *Asian Journal of Marketing, 5*(1), 41–58.

Wittink, D. A., Krishnamurthi, L., & Nutter, J. B. (1982). Comparing derived importance weights across attributes. *Journal of Consumer Research, 8*(4), 471–474.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western Cengage Learning.

译名对照表

aliased	别名的
analysis of variance, ANOVA	方差分析
anchoring effects	锚定效应
anchoring vigettnes approach	锚定情境法
blocks, blocking	区组(块)
bouncing beta problem	跳跃 β 问题
censored response	应答删失
choice experiments, CEs	选择性实验
conditional logit, CL	有条件的 logit(模型)
confounded	混淆的
conjoint analysis, CJA	联合分析法
construct validity	构念效度
cross-elasticity	交叉弹性
cross-validate	交叉校验
decks	层块
design resolution	设计分辨度
Fisher information matrix, FIM	Fisher 信息矩阵
fixed effect, FE	固定效应
fractional factorial	部分因子
fractionalized samples	分部样本
full factorial	全部的因子
halo effect	晕轮效应
heteroscedasticity	方差不齐性
hierarchical linear models, HLM	分层线性模型
homoscedasticity	方差齐性
hurdle models	跨栏模型
independence of irrelevant alternative, IIA	无关备选方案的独立性假设
interval-scaled variable	区间标度变量
likelihood ratio, LR	似然比
long format	长格式
multiple segment factorial design	多部分因子虚拟情境设计

number-matching technique	数字匹配技术
number-of-levels effects	水平数量带来的影响
orthogonal	正交
parameters of interest	增益参数
partial confounding	部分混淆
priors	先验
psychological realism	心理现实主义
random effect, RE	随机效应
random intercept models	随机截距模型
random intercept, RI	随机截距
random slope models	随机斜率模型
random utility theory, RUT	随机效用理论
semi-partial	半偏
sets	集合
simplifying heuristics	简化启发式
social desirability bias, SDB	社会赞许性偏见
split-ballot design	分签一投票设计
statistical power calculations	统计检验力计算
status quo bias	现状偏见
trade-offs	权衡取舍
vignette population	总体情境
vignette	虚拟情境法
wide format	宽格式
wildcards	通配符

译后记:以虚拟接近真实

当整沓校样摆在案头的时候,我不禁感叹起自己 2017 年不知天高地厚地申请翻译这本书的情形。那时刚写完全国教育科学规划项目的申请书,我在设计《新型教育舆情的政策回应》的课题时,就深感如果只是用传统的问卷调查直接搜集大众对教育舆情和教育政策看法,一定存在诸多明显的局限。

如何突破?如何在传统的研究方法上进行可操作的创新?顺藤摸瓜,我找到了因素调查这个方法,由此也就发现了 SAGE 出版社的这本《因素调查实验》。机缘巧合,我向格致出版社申请翻译这本书,并顺利通过立项。

在奥斯普格与欣茨的这本《因素调查实验》中,他们主要围绕"虚拟情境法"设计问卷中的实验刺激与干预,通过将虚拟情境随机分配给控制组与实验组的方式来确认因果关系。虚拟情境法是我们刻意选择的中文译名。该词的英文是 vignette,原指一段能清晰展示任务特征和局势的短文,20 世纪 50 年代就被应用于问卷调查研究。为了突出这种根据实验干预需要而拟定的"虚拟性",我们刻意在情境之前加上

“虚拟”一词。但是这个方法最吊诡和最具价值的地方就在于以虚拟接近社会的真实。

目前随着研究方法的普及，中国社会科学研究中大量使用问卷调查的量化研究方法。确切地说，问卷只是实证数据的搜集手段和方法。研究风潮一起，众多研究者对研究方法采用的是“依样画葫芦”的简单“拿来主义”，并未真正考虑研究工具、数据搜集方法、数据分析方式与研究问题和研究目的的匹配。至少在我熟悉的教育领域中，问卷调查已经成为量化实证研究的主力军，但问卷调查结果多大程度反映了真实的世界、真实的观念、真实的问题，却少有人质疑。

在强调实证研究的时代，不仅要使用研究方法，更要考虑使用什么研究方法、如何使用更多样有效的方法才能达致我们做真研究的目的。我相信读者们在通读全书以后，一定会了解到因素调查实验为什么比简单的问卷调查更能有效地探查真实世界中的复杂因果关系。除了因素调查实验，还有更多社会科学研究设计方法可以与问卷相结合进行复杂和深层的因果分析。我也是通过翻译此书更体会到了在看似复杂的数据模型之外，更应该从研究设计角度构思研究，不为数据之“术”所困。

在主动寻求翻译的时候，我知道学术翻译不算任何工作量或岗位贡献，但从个人学习收获和推动中国社会科学研究方法与实证研究多样性的角度，我还是很感恩有机会翻译本书。由于个人工作任务的突然调整与书中的技术细节等原因，这本书在荣佳妮老师、叶晓阳博士和王哲博士的帮助下，终于在 2020 年疫情期间完成了。这三位都熟练使用量化方法，尤其是叶晓阳博士与王哲博士专精于教育中的各种实验

研究，发表过不少应用类似方法的研究论文。由于知识和经验的局限，本文的错误均由本人承担，特别欢迎各位读者和专家就有关错误联系我，我将与出版社一起以合适的方式修正错误，为中国社会科学研究贡献微薄之力。

陈霜叶

2021年2月26日于水思悦斋

格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用 logistic 回归分析（第二版）
14. logit 与 probit：次序模型和多类别模型
15. 定序因变量的 logistic 回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic 回归入门
39. 解释概率模型：Logit、Probit 以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL 导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic 回归中的交互效应
46. 应用回归导论
47. 档案数据处理：生活经历研究
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践
51. 指数随机图模型导论
52. 对数线性模型的关联图和多重图
53. 非递归模型：内生性、互反关系与反馈环路
54. 潜类别尺度分析
55. 合并时间序列分析
56. 自助法：一种统计推断的非参数估计法
57. 评分加总量表构建导论
58. 分析制图与地理数据库
59. 应用人口学概论：数据来源与估计技术
60. 多元广义线性模型
61. 时间序列分析：回归技术（第二版）
62. 事件史和生存分析（第二版）
63. 样条回归模型
64. 定序题项回答理论：莫坎量表分析
65. LISREL 方法：多元回归中的交互作用
66. 蒙特卡罗模拟
67. 潜类别分析
68. 内容分析法导论（第二版）
69. 贝叶斯统计推断
70. 因素调查实验

[General Information]
书名=因素调查实验
页数=187
SS号=14916874